HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Screening of bacteriophage encoded toxic proteins with an NGS based assay

Jutta Kasurinen

Master's thesis

Master's programme in Microbiology and Microbial biotechnology

June 2020

Tiedekunta – Fakultet – Faculty	Koulutusohjelma – Utbildningsprogram – Degree Programme
Faculty of Agriculture and Forestry	Master's programme in Microbiology and Microbial Biotechnology

Tekijä – Författare – Author
Jutta Kasurinen

Työn nimi – Arbetets titel – Title
Screening of bacteriophage encoded toxic proteins with an NGS based assay

Työn laji – Arbetets art – Level	Aika – Datum – Month and year	Sivumäärä – Sidoantal – Number of pages
Master's thesis	June 2020	44

Tiivistelmä – Referat – Abstract
The ever-increasing spread of antibiotic resistant bacteria creates a constant demand for new sources for antimicrobial drugs. Phages are a natural source for antibacterial proteins, but also produce a variety of unknown compounds, referred to as "hypothetical proteins of unknown function" (HPUF). HPUFs usually consist of structural proteins, but also small polypeptides that inhibit bacterial growth during infection. These peptides could be utilized in the discovery of new antimicrobial molecules. However, the current methods used for the screening of such proteins are time consuming and unreliable, making this a fairly unpopular option to utilize.

In this study, a new NGS (Next Generation Sequencing) based assay for the screening of phage derived bacteriotoxic proteins was developed and tested by performing two separate experiments together with a previously used plating assay as a comparative method. A preliminary experiment was performed as a proof of principle, with five known toxic and five non-toxic genes. After this, the methods were compared by screening 23 previously identified HPUF genes of phage fHy-Eco03. In the plating assay genes were screened individually by observing growth of bacterial transformants upon gene expression. In the NGS assay genes we screened simultaneously by transforming them to *E. coli* cells as a pooled sample. Results were obtained with bioinformatics. Toxic genes were expected to be identified through a decrease in sequence read amount, as a consequence of bacterial growth inhibition.

In the pre-experiment a difference between toxic and non- toxic proteins was not observed. The results between the NGS and plating assay in the screening of phage fHy-Eco03 genes, were similar and resulted in the identification of one toxic protein. The inconsistent results are probably an outcome of *lac* promoter repression by glucose supplementation, thus only highly toxic genes show an inhibitory effect. Despite this the NGS assay outperformed the plating assay in both accuracy and efficiency. The NGS assay has high potential to be used as a screening assay for phage derived toxic genes, however further optimization and validation is required, by firstly selecting compatible media and secondly by re- testing with different phages and host bacteria.

Avainsanat – Nyckelord – Keywords
Bacteriophage, antibiotic resistance, hypothetical proteins of unknown function, Next Generation Sequencing

Ohjaaja tai ohjaajat – Handledare – Supervisor or supervisors
Maria Pajunen, Mikael Skurnik

Säilytyspaikka – Förvaringsställe – Where deposited
HELDA – Helsingin yliopiston digitaalinen arkisto

Tiedekunta – Fakultet – Faculty	Koulutusohjelma – Utbildningsprogram – Degree Programme
Maatalous- ja Metsätieteellinen tiedekunta	Mikrobiologian ja Mikrobibiotekniikan maisteriohjelma

Tekijä – Författare – Author
Jutta Kasurinen

Työn nimi – Arbetets titel – Title
Bakteriofagien tuottamien toksisten proteiinin identifioiminen NGS:ään perustuvalla menetelmällä

Työn laji – Arbetets art – Level	Aika – Datum – Month and year	Sivumäärä – Sidoantal – Number of pages
Pro-gradu tutkielma	Kesäkuu 2020	44

Tiivistelmä – Referat – Abstract

Antibioottiresistenssin jatkuva leviäminen on lisännyt tarvetta uusille antibiooteille. Bakteriofagit tuottavat antibakteerisia proteiineja, mutta myös niin sanottuja hypoteettisia proteiineja (HPUF) infektiosyklin aikana. Näiden proteiinien joukossa on sekä rakenteellisia, että pieniä antibakteerisia polypeptideitä, jotka estävät bakteerisolun toimintaa infektion aikana. Kyseiset peptidit ovat potentiaalinen lähde uusille mikrobilääkkeille. Näiden molekyylien tunnistamiseen perinteisesti käytetyt menetelmät ovat kuitenkin erittäin työläitä ja epäluotettavia, eivätkä siten kannusta kyseisten proteiinien seulontaa uusien lääkkeiden etsinnässä.

Tässä tutkielmassa, sekä kehiteltiin, että testattiin uuden sukupolven sekvensointiin (NGS, next generation sequencing) perustuvaa menetelmää bakteriofagien tuottamien toksisten proteiinien identifioimiseen. Tutkimus suoritettiin toteuttamalla kaksi erillistä koetta, joissa vertailtiin NGS menetelmää aiemmin käytettyyn maljausmenetelmään. NGS menetelmän toimivuutta testattiin esikokeella, jossa seulottiin viisi tunnettua toksista ja ei-toksista geeniä. Tämän jälkeen seulottiin 23 bakteriofagi fHy-Eco03 tuottamaa hypoteettista proteiinia. Maljausmenetelmässä proteiinit seulottiin yksitellen seuraamalla transformanttien kasvua proteiinieksprression jälkeen. NGS menetelmässä proteiineja seulottiin analysoimalla geenisekvenssejä yhdistetyistä plasmidinäytteistä. Toksisten proteiinien identifioiminen perustuu niitä koodaavien geenien alhaisempaan havaittuun sekvenssimäärään isäntäsoluun kohdistuvien toksisten ominaisuuksien seurauksena.

Esikokeen perusteella NGS menetelmällä ei havaittu selkeitä eroja toksisten ja ei-toksisten proteiinien välillä. Bakteriofagi fHy-Eco03 seulonnassa NGS- ja maljausmenetelmä-tulokset olivat hyvin samankaltaisia. Seulonnassa identifioitiin yksi toksinen proteiini. Tuloksiin on oletettavasti vaikuttanut plasmidissa sijaitsevan *lac*-promoottorin repressointi, elatusaineen glukoosisuplementin johdosta. Täten vain erittäin toksiset geenit inhiboivat bakteerikasvua. Huolimatta kyseisistä ongelmista, NGS menetelmä osoittautui erittäin potentiaaliseksi bakteriotoksisten proteiinien identifioimisessa, sillä se oli, sekä tehokkaampi, että tarkempi maljausmenetelmään verrattuna. Menetelmän käyttöönotto vaatii kuitenkin validoimista ja optimoimista, esimerkiksi kasvualustan suhteen. Lisätesaus muilla bakteriofaageilla sekä isäntäsoluilla on myös tarpeen.

Avainsanat – Nyckelord – Keywords
Bakteriofagi, antibioottiresistenssi, hypoteettiset proteiinit, uuden sukupolven sekvensointi

Ohjaaja tai ohjaajat – Handledare – Supervisor or supervisors
Maria Pajunen, Mikael Skurnik

Säilytyspaikka – Förvaringsställe – Where deposited
HELDA - Digital Repository of the University of Helsinki

INTRODUCTION

Ever since the discovery of first antibiotics, the unrestricted access and irresponsible use has resulted in the ever-increasing emergence of multidrug-resistant (MDR, multiple drug resistance) bacterial strains resistant to nearly all currently available antimicrobial drugs (Cooper & Shlaes, 2011; Munita & Arias, 2016). During the last 30 years, the abundance of methicillin resistant *Staphylococcus aureus* (MRSA) (Rice, 2006) and extended-spectrum β-lactamase (ESBL) (Bush, 2010) strains has increased drastically, with some strains reported to be resistant even to the so-called last resort drugs such as polymyxin B and colistin (Jorge et al., 2017). The development and release of new antibiotics is highly demanding, tedious and low in revenue, hence only two new classes of antibiotics have been released in the last 20 years (Tacconelli et al., 2018; Nielsen et al., 2019). Antibiotics have for long been developed by enhancing the existing ones by modifications, creating a high demand for new sources of antimicrobials (Laxminarayan et al., 2013; Fernandes & Martens 2017; Tacconelli et al., 2018).

Bacteriophages, viruses that infect bacteria, are considered as one, if not the most abundant organisms on earth, with an estimated global population size of 10^{31} (Hendrix et al., 1999) Approximately 90 % of viral sequences are not present in any databases (Hendrix, 2002; Ackermann, 2011). Since their discovery by FW Twort in 1915 (Salmond & Fineran, 2015), bacteriophages and phage derived proteins have been researched and used as antibacterials by the food industry (Endersen et al., 2014) and as therapeutics in health care (Karthik et al., 2014; Schmelcher & Loessner 2016). Viriolysins are antibacterial proteins with bacteriolytic and bacteriostatic properties. These proteins, produced in the early and late stages of viral infection are involved in the degradation of peptidoglycan, a rigid polymer forming the cell wall of bacteria (Parisien et al., 2008). The most known viriolysins are endolysins, holins and virion-associated peptidoglycan hydrolases (VAPGH). Other known, but less studied phage derived antibacterial proteins are holin-type lytic factors and phage tail complexes associated with the inhibition of peptidoglycan biosynthesis and cell wall degradation, respectively (Young, 2002; Parisien et al., 2008; Schmelcher & Loessner, 2016).

Bacteriophages also produce vast amounts of compounds whose functions are unknown because of the lack of reference in sequence databases. These compounds are referred to as "hypothetical proteins of unknown function" (HPUF). Hypothetical proteins usually consist of structural proteins but also small polypeptides that are produced during the infection to affect or inhibit the cellular mechanism of the host that could otherwise inhibit the phage infection cycle (Van Den Bossche et al., 2014). These toxic proteins could be the source of new antimicrobial molecules

as indicated by previous research (Liu et al., 2014; Shibayama & Dabbs, 2011; Mohanraj et al., 2019; Singh et al., 2019; Spruit et al., 2020).

Toxic HPUFs have previously been screened individually by utilizing affinity purifications combined with mass spectrometry (AP−MS) during infection (Van Den Bossche et al., 2014) and by observing growth inhibition upon HPUF expression (Liu et al., 2014; Shibayama & Dabbs, 2011; Mohanraj et al., 2019; Singh et al., 2019; Spruit et al., 2020). Two of these methods used shotgun cloning of genomic fragments and identified toxic ORFs (open reading frames) by comparing transformation efficiencies against plasmid controls (Shibayama & Dabbs, 2011) or by measuring growth upon the expression of inducible vectors (Singh et al., 2019). The disadvantage of these methods is the plausible overlooking of toxic ORFs, either because of gene fragmentation or incorrect orientation during cloning. In another method (Liu et al., 2004), a total of 964 genes of 27 *Staphylococcus aureus* phages were screened by first excluding proteins with known or predicted toxic or structural functions before cloning the genes into inducible expression vectors. This minimizes time and resources as surplus genes with structural or other functions are not screened. Plating assay described by Mohanraj et al., (2019) and Spruit et al., (2020), is an adaptation of the assays described above. With this assay HPUFs of phages ϕR1-RT (Mohanraj et al., 2019) and fHe-Kpn01 (Spruit et al., 2020) were discovered, by first identifying true hypotheticals with LC-MS/MS analysis and annotation. Bacteriotoxic properties were screened by comparing transformation efficiencies against non-toxic control genes. The toxicity of obtained hits was then confirmed by cloning the candidate genes into an arabinose-inducible reporter plasmid and observing bacterial growth during arabinose induced expression.

The plating assay-based screening and the other methods described above, are, however, due to many different difficult-to-standardize steps prone to high variation and require multiple replicates and repetitions to achieve statistically significant results. This consequently reduces the economical application of these assays. The amount of time and resources required to produce valid results increases drastically with the number of genes, limiting the amount that can be screened at a time to ensure somewhat equivalent conditions for all the samples. These issues with reliability are most probably the outcome of variation in quality of the DNA, experimental conditions and the manual lab work, including pipetting, plating and false positives from undigested or incorrectly ligated plasmids. In addition, it is important to note that electroporation is not an ideal method for this type of quantitative work since the conditions cannot be entirely controlled leading to major differences between replicate transformations.

The development of Next generation sequencing (NGS) has enabled simultaneous sequencing of millions of DNA molecules. This has created the possibility to do targeted sequence

analysis from vast amounts of pooled data (Von Bubnoff, 2008). NGS based screening could therefore be used to overcome the obstacles present in the methods described above, as plasmids could be transformed as a pooled sample and analyzed by sequencing and bioinformatics. The principle behind this method is that the sequence read coverages of genes should be in relation to the transformants carrying them. Therefore, the detected sequences of toxic genes are expected to decrease as a consequence of bacterial growth inhibition.

Phage fHy-Eco03 is a dwarf Myovirus, infecting *Escherichia coli*, a gram-negative bacterium of the *Enterobacteriaceae* family, with a high clinical relevance as several MDR strains have been identified globally (Petty et al., 2014; Rogers et al., 2011). Phage fHy-Eco03, isolated from a hospital sewage sample (Hyvinkää), has been sequenced and annotated previously in the Skurnik group (Yersinia and bacteriophage research laboratory, University of Helsinki). It has a genome of 54 kb containing 80 predicted genes among which 32 HPUF-encoding genes were identified based on annotation and identification of phage-particle associated proteins (PPAPs) by LC-MS/MS (liquid chromatography–mass spectrometry) proteomics. The host range of fHy-Eco03 was tested against 50 *E. coli* strains, of which the phage was capable of infecting two strains, both resistant to ampicillin, trimethoprim, sulfonamides and tobramycin. (Wicklund, 2014).

In this study an NGS based screening assay was developed and tested by performing two separate experiments. A preliminary experiment, performed as a proof of principle, was carried out using five genes encoding known toxic and non-toxic proteins of phages ϕR1-RT (Moharanj et al., 2019), fHe-Kpn01 (Spruit et al., 2020) and T4 (Ruckman et al., 1994). After this 23 previously identified HPUF encoding genes of phage fHy-Eco03 were screened for toxic ones in a comparative study with the NGS-approach and the plating assay described in Mohanraj et al (2019).

MATERIALS AND METHODS

Bacterial strains and culture conditions

Bacterial strains and bacteriophages used in this study are listed in Table 1. Bacteria were grown in LB broth or on solid LB agar (1.5% (w/v) agar) (LENNOX, Difco) supplemented with 100 µg/ml ampicillin (Sigma-Aldrich, St. Louis, MO, USA) unless mentioned otherwise. S.O.B medium (Inoue et al., 1990) was used for the preparation of electrocompetent cells (Lamberg et al., 2002) and S.O.C medium (Dower et al., 1988) was used as a growth medium in transformations. M9t minimal medium (3.4 mM Na_2HPO_4, 2.2 mM KH_2PO_4, 0.94 mM NH_4Cl, 0.86 mM NaCl, 0.2% (w/v) tryptone, 2.0 mM $MgSO_4$, 0.10 mM $CaCl_2$ and 3.0×10^{-3} mM vitamin B1) was used as a growth medium for transformed DH5α cells. Bacteria were grown under aerobic conditions at 35°C or 37°C, either using a slow 14 rpm rotation or 200 rpm shaker for 1-24 h according to method.

Table 1. Bacterial strains and bacteriophages utilized in this study

	Strain	Purpose	Storage code	Source
E. coli	DH10B	Preparation of electrocompetent cells	#2689	Skurnik lab strain collection
	DH5α		#6152	
	DH10B/pU11L4	Source of plasmid pU11L4	#6664	
Bacteriophage	ϕ R1-RT	Template for amplification of ϕR1-RT genes	#3	Skurnik lab strain collection
	T4	Template for amplification of T4 genes	#67	
	fHe-Kpn01	Template for amplification of fHe-Kpn01 genes		
	fHy-Eco03	Template for amplification of fHy-Eco03 genes		

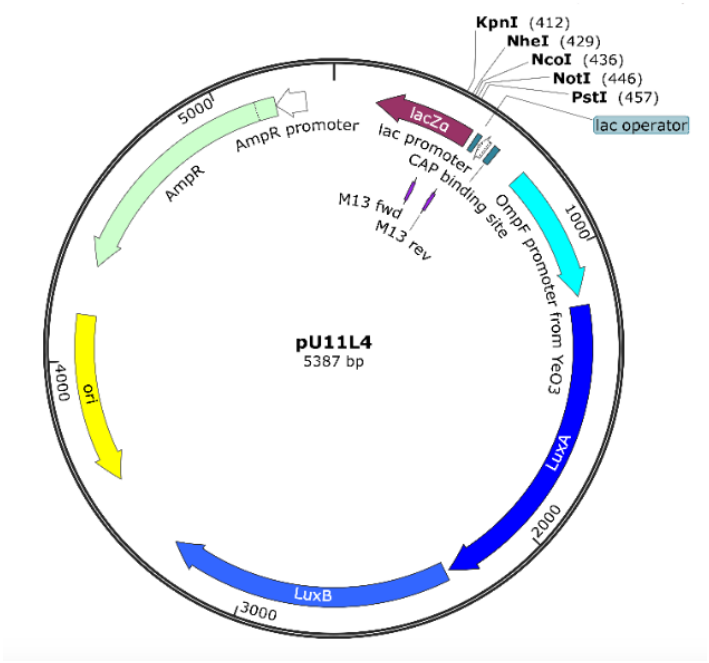

Figure 1. Map of plasmid pU11L4. The Figure was created using SnapGene viewer (GSl Biotech; snapgene.com). (Mohanraj et al., 2019; Supplementary material)

Recombinant DNA methods

The plasmid pU11L4 (Figure 1), used as the cloning vector in this study, was isolated from o/n grown cells with a commercial Nucleobond Xtra Midi kit (MACHEREY-NAGEL, Germany) according to the protocol for high-copy number plasmids. Plasmid pU11L4 was double-digested with restriction enzymes *Not*I / *Nco*I and *Nhe*I/ *Not*I (Thermo Fisher Scientific, USA) accordingly in 10× FD buffer (Thermo Fisher Scientific, USA). The digestions were incubated overnight at 37°C and dephosphorylated with FastAP™ Thermosensitive Alkaline Phosphatase (Thermo Fisher Scientific, USA) by incubating at 37°C for 30 min. The plasmid digestions were checked by agarose gel electrophoresis, and approved digestions were then run in preparative agarose gel (SeaPlaque GTG agarose BMA) at 50V for 3 h, with GeneRuler™ 1 kb DNA Ladder (Fermentas) as a size marker. The linearized plasmid of correct size was excised from the gel under preparative UV illumination and purified with NucleoSpin® Gel and PCR Clean-up kit (MACHEREY-NAGEL, Germany) according to instructions.

Toxic and non- toxic control genes and fHy-Eco03 HPUF genes were amplified by PCR with primers containing the appropriate restriction sites at the 5’ end of each primer (Table 2). The PCR reactions were performed in 50 µl volumes containing 500 nM of each primer (Metabion,

Germany) and 0.02 U/µl of Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific, USA). The PCR program consisted of a 3 min denaturation at 98°C and then 30 cycles of denaturation at 98°C for 10 sec, annealing at 58°C for 30 sec and extension at 72 °C for 60 sec. The program ended with a 10 min extension step at 72°C and infinite hold at 4°C. Five µl of the PCR products were run in 1% agarose gel (SeaKem LE agarose gel, BMA) at 120 V for 45 min. The amplified control and fHy-Eco03 HPUF genes were digested with appropriate restriction enzymes, in 10× FD buffer by incubating overnight at 37°C. All PCR products and restriction-digested genes were purified with NucleoSpin® Gel and PCR Clean-up kit (MACHEREY-NAGEL, Germany) according to instructions.

Control genes and fHy-Eco03 HPUF genes were cloned accordingly to the *Not*I and *Nco*I or *Not*I and *Nhe*I restriction sites of linearized plasmid pU11L4. For each ligation, the digested Insert and plasmid was mixed in a molar ratio of 3:1 with 3 U of T4 DNA ligase (NEB, USA and 10× T4 Ligation Buffer (NEB, UK). The ligations were incubated at RT for 45 min, then o/n at 16°C, heat inactivated at 65°C for 25 min and stored at -20°C. Electroporation was performed in 0.2 cm cuvettes by combining 45 µl of electrocompetent *E. coli* DH10B cells with 1 µl of DNA. The pulse was given with a Gene Pulser™ apparatus (Bio-Rad Laboratories, USA), with the following settings: 200 Ω, 25 uF and 2.5 kV. Transformed cells were grown in 950 µl of S.O.C medium at 35°C for 1 h in slow rotation before plating 50 µl on LB Ampicillin plates. Plates were incubated o/n at 37°C. Transformations of fHy-Eco03 HPUFs were done in batches of 4 to 6 samples with *g178* of phage ϕR1-RT as a non- toxic control. The relative CFUs were determined from triplicate plating's of two biological replicates as a fraction of the non-toxic control. The CFUs of control genes were determined from triplicate plating's of two biological replicates.

The NGS assay was carried out with two biological replicates of pooled ligation mixtures. From each mixture two replicate transformations were performed. All individual fHy-Eco03 HPUF ligations were pooled together (no control genes) and the ligation mixture was purified as described before, to concentrate the sample and get rid of salts that might affect the transformation efficiency. A sample of the pooled ligation mixture was withdrawn for NGS analysis. Electroporation was performed as described previously. After electroporation the cells were grown for 1 h in S.O.C at 37°C in slow rotation before plating all of the inoculation to LB Ampicillin plates. Colonies from o/n grown plates were collected with S.O.C medium and grown in a 10-fold dilution supplemented with ampicillin (100 µg/ml) for 3 hours at 180 rpm at 37°C. After this, cells were pelleted (3000 RCF, 10 min, 4°C) and plasmid isolation was carried out with Nucleobond Xtra Midi kit (MACHEREY-NAGEL, Germany) according to instructions.

Table 2. Primers used to amplify the toxic and non-toxic control genes by PCR

Non- toxic genes

Template	Primer	size (bp)	Primer sequence (5' – 3')	Restriction site
ϕ R1-RT	Gp119F	591	GCAGCGGCCGCATGAAAACGTATAAAGAATTTTTG	*Not*I
	Gp119R		GGTCCATGGTTAACGAACGTTAGTGCCA	*Nco*I
	Gp121F	273	GCAGCGGCCGCATGAAAACCTATAATGAATTTATC	*Not*I
	Gp121R		GGTCCATGGTTAGGAAGCTTTTTTAAGC	*Nco*I
	Gp150F	1728	GCAGCGGCCGCATGATTAAAGTTAATGAGC	*Not*I
	Gp150R		GGTGCTAGCCTATCCAATATCAATTCGTGAA	*Nhe*I
	Gp178F	1275	GCAGCGGCCGCATGAGCAATATTAACCAGC	*Not*I
	Gp178R		GGTCCATGGTTATCCTGCTATTAGTTTAGG	*Nco*I
	Gp246F	939	GCAGCGGCCGCATGTCTTTAAATGAAATG	*Not*I
	Gp246R		GGTCCATGGTTAAAAATCATTTGCATG	*Nco*I

Toxic genes

Template	Primer	size	Primer sequence (5' – 3')	restriction site
fHy-Kpn01	Gp10F	210	GCAGCGGCCGCATGATTAAGTACGATGTATACAAG	*Not*I
	Gp10R		GGTCCATGGCTACTGTGAGCATAGGCTG	*Nco*I
	Gp22F	549	GCAGCGGCCGCATGATTGACAGAGAAGAGATAC	*Not*I
	Gp22R		GGTCCATGGTTAATATGCATCACGCACC	*Nco*I
	Gp38F	441	GCAGCGGCCGCATGAAGTTAAAACACACTAGTAA	*Not*I
	Gp38R		GGTCCATGGTCATTCTGACCTCACTAAATG	*Nco*I
ϕR1-RT	Gp137F	552	GCAGCGGCCGCATGAAAATTGCTGAACTAAT	*Not*I
	Gp137R		GGTCCATGGCTATGAGGACTTAGAAATTGT	*Nco*I
T4	regBF	491	GATCGCGGCCGCCCATGACTATCAATACAGAAG	*Not*I
	regBR		GGCCGCTAGCCTTACCTCATTGAGTTTTAATTAC	*Nhe*I

Sequence analysis

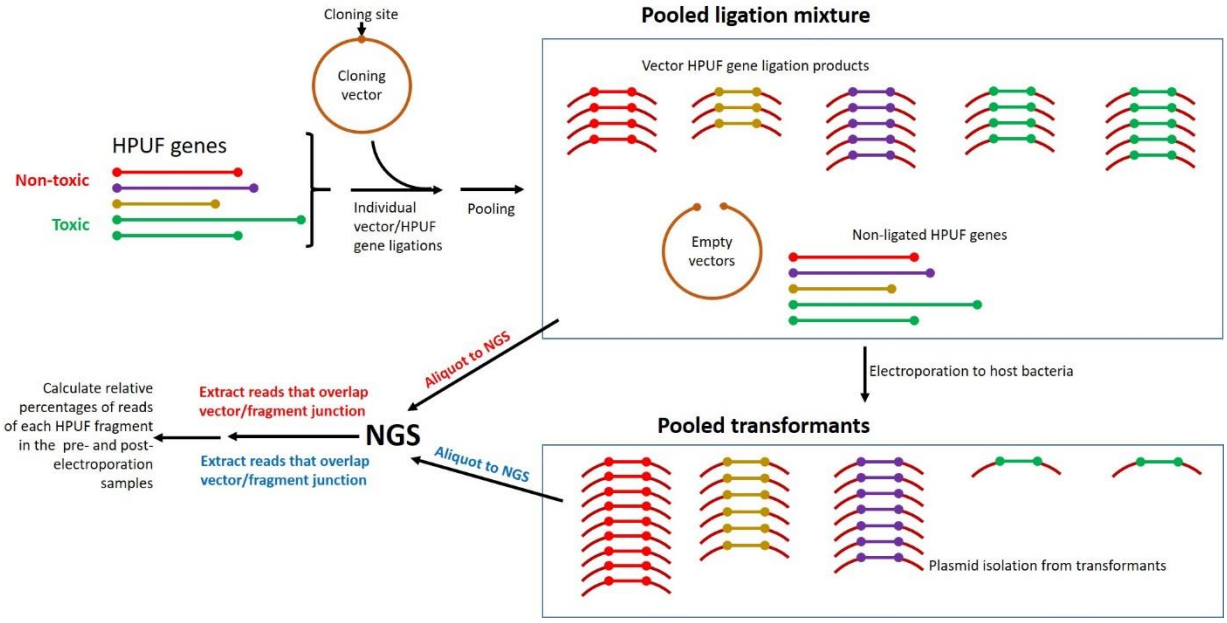

Figure 2A. Schematic illustration of the NGS-based approach to identify phage HPUFs encoding toxic proteins.

The NGS-based screening approach is outlined in Figure 2A. The ligation mixture and plasmid DNA samples were sequenced using the 150 bp paired end protocol in the Illumina HiSeq platform at NovoGene (UK). Sequences were analyzed using the Puhti computer environment at CSC (the Finnish Centre for Scientific Computing). The workflow of the bioinformatics is described in detail in the supplementary material S6. In order to obtain information about the amount of correctly ligated genes, four predicted sequence fragments over the ligation joints per gene were aligned against the NGS raw read data. Each fragment, presented in Figure 2B, contains a restriction site and approximately 15 - 25 nucleotides from the plasmid and insert, including either a start (atg) or a stop codon of the gene. The fragments are derived from both leading and lagging strands of the plasmid, resulting in complementary pairs of forward and reverse sequences. The combined amount of reads from these sequences represents the abundance of each HPUF in a sample. Relative percentages were determined for each HPUF in the pre- and post-electroporation samples. The relative percentages of HPUFs in the post-electroporation (transformant plasmid) sample were then divided with the corresponding percentages in the pre-electroporation (ligation mixture) sample, thus producing a ratio of relative reads for each gene. This ratio describes the change in read amounts between the pre- and post-electroporation samples. Protein sequence alignments were performed against protein sequence

databases with BLASTx (Altschul et al., 1990) and the protein structures were modeled using Phyre2 program (Kelley et al., 2015).

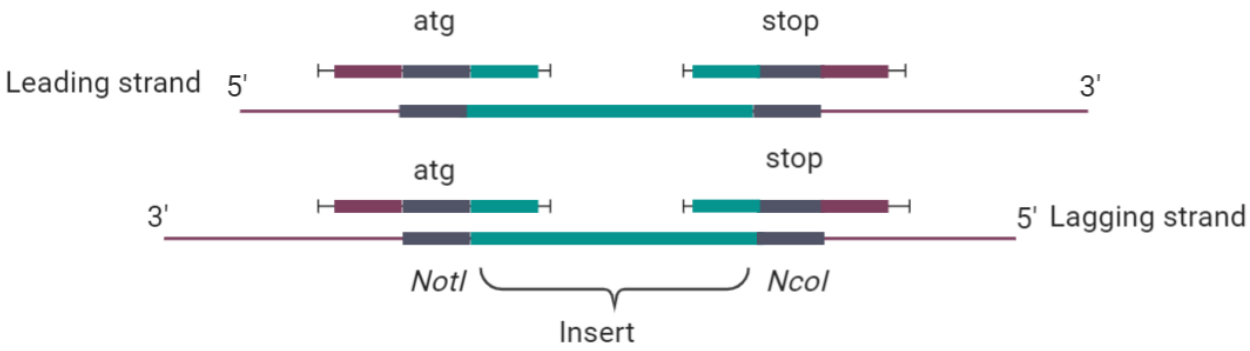

Figure 2B. An illustration of the four sequence fragment constructs used in alignments against plasmid sequences. The image was created in BioRender.com.

Growth curve analysis

Growth curve analysis with Bioscreen was used to confirm the toxicities of selected HPUF genes. Genes were amplified with primers containing pBAD30 compatible RE-sequences at the 5' ends of the primers. The amplified genes were purified and cloned into an inducible pBAD30 vector using the *Kpn*I and *Xba*I or *Sph*I restrictions sites accordingly. The plasmids were transformed into electrocompetent *E. coli* DH5α cells as described earlier. Transformant colonies and colonies containing pBAD30 vectors with control gene *g137* (toxic) and *g150* (non-toxic) insertions and empty plasmid pBAD30 were picked from plates and grown o/n at 37°C at 180 rpm in LB medium supplemented with glucose (0.2% w/v) and ampicillin (100 µg/ml). Afterwards the cells were pelleted and resuspended in M9t media and 10 µl of cells were inoculated in 1 ml of M9t (100 µg/ml Amp) supplemented with either glucose (0.2% w/v) or arabinose (0.2% w/v). Bacterial inoculations were pipetted to Bioscreen Honeycomb plates and the ODs (600 nm) were measured with Bioscreen C MBR (Oy Growth Curves Ab Ltd, Helsinki, Finland) at an hourly rate for 20 hours. The average ODs were determined from triplicate wells of three biological replicate cultures of each gene. Successful cloning was also confirmed by isolating and sequencing plasmids from the transformed DH5α cells. Sanger sequencing was performed at the Finnish Institute for Molecular Medicine (FIMM).

RESULTS

Preliminary experiment

In order to determine whether the NGS assay could function as a reliable screening method for the detection of phage encoded toxic proteins, a pre-experiment was performed with five known toxic and non-toxic genes (Table 2). Ligations and toxicities of the control genes were first confirmed individually with the plating assay. The transformation efficiencies of the non-toxic controls were from two to several orders of magnitude higher than the toxic controls as expected, varying from an average of ~280 (*g150*) CFUs to ~580 (g178) CFUs per 50 µl of cells. The CFUs of toxic controls varied from a few colonies (*regB*) to ~100 (*g137*) (Figure 3).

In the NGS assay results are presented as a ratio of relative reads that describes the change in gene sequence read percentages between the ligation mixture and transformant plasmid sample. The NGS assay results presented in Figure 3 (in detail in supplementary Table S3) showed no clear differences between toxic and non-toxic controls. The relative amounts of reads from toxic genes *g22*, *g10* and *g83* were 97% (0.003) and 85% (0.015) lower and 50% (1.5) higher in the transformant, than in the ligation mixture sample respectively. In comparison, the relative ratios of non-toxic controls varied between 0.19 (*g150)* and 5.6 (*g121*), representing 80% and 500% higher relative amounts in transformant, than in the ligation mixture derived reads. Toxic control gene *g137* differed substantially from all genes with a ratio of 27.8. Toxic gene *regB* fragments were not found to be present in any of the NGS samples.

The proportions of sequence read coverages in the ligation mixture and transformant plasmid samples are presented in Table 3. The relative amounts of reads in the ligation mixture were quite disproportionate varying from the lowest of ~0.5% and ~0.9% of *g137* and *g10* respectively to 23% of *g246*. The amount of reads in the transformant samples varied in average from the lowest of 0.07% and 0.13% of toxic genes *g22* and *g10* to 26.8% of toxic gene *g38*. The amount of *g137* reads were 13.6% thus explaining the high relative ratio (27.8). Based on the results the only toxic genes presenting bacteriotoxic effects were *g10* and *g22*.

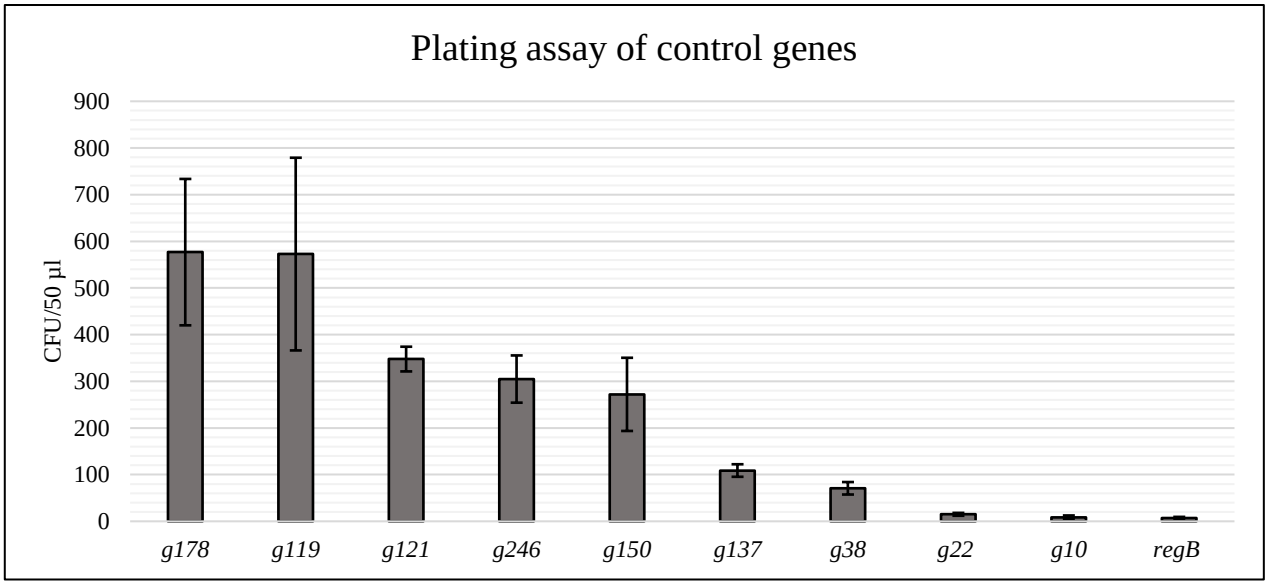

Figure 3. Preliminary experiment results of the plating assay of control genes. The results are presented as CFUs of toxic and non-toxic control genes per 50 µl of cells plated after electroporation. Genes *g178, g119, g121, g246* and *g150* are non-toxic and genes *g137, g38, g22, g10* and *regB* are toxic. Bars indicate mean ± SD across triplicate plating's of two biological replicates of each gene.

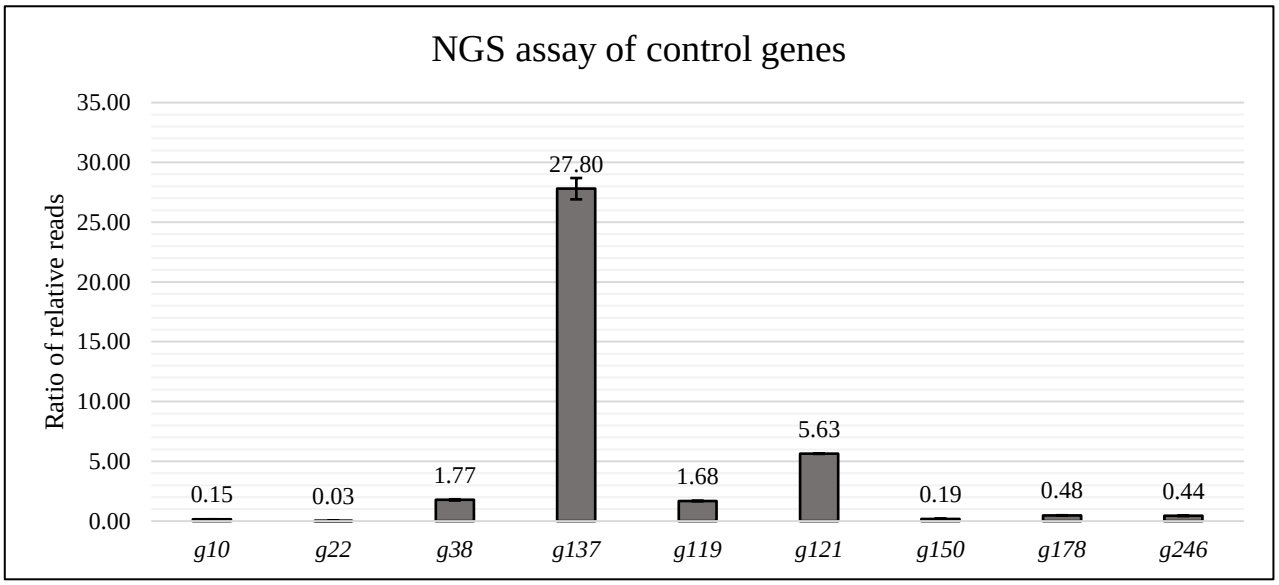

Figure 4. NGS assay results of the preliminary experiment. The results are shown as ratios of the relative reads between the ligation mixture and the transformant plasmid samples. Genes *g178, g119, g121, g246* and *g150* are non-toxic and genes *g137, g38, g22* and *g10* are toxic. The bars stand for mean ± SD between two replicate transformations.

Table 3. NGS assay results of the preliminary experiment. Relative amounts (%) of control gene ligation products in the ligation mixture and transformant plasmid samples. The transformant plasmid sample results are presented as averages and standard deviations (SD) between two replicate transformations.

	Relative amounts (%)		
Gene	Ligation mix	Transformant sample	SD
g10	0.86	0.13	0.001
g22	3.10	0.07	0.019
g38	14.84	26.83	0.513
g137	0.48	13.16	0.305
g119	11.70	19.53	0.183
g121	2.79	15.74	0.002
g150	21.38	3.81	0.194
g178	21.74	10.58	0.155
g246	23.11	10.17	0.034

Screening of fHy-Eco03 HPUF genes

Plating assay

Putative toxicities presented as fractions of non-toxic control gene *g178* are shown in a descending order in Figure 5. Average fractions varied from the lowest of 0.16 of gene *g05* to 1.71 of gene *g48*, representing approximately 16% and 170% of gene *g178*. In addition to gene *g05*, values of nine other genes with the lowest fractions varied from 0.33 (*g52*) to 0.68 (*g01*) representing approximately 30% and 70% of *g178* (Table 5). The results between replicate genes varied at highest over two-fold (*g77*) as can be seen from Figure 5 and in detail in the supplementary Table S4. For example, batch to batch variation of non-toxic control *g178* ranged from approximately 300 to 700 CFUs (Table S4). The standard deviations between the two replications varied from the lowest of 0.003 (*g05*) to 0.505 (*g76*) with a majority deviating between 0.100 and 0.300 (Table 5). The CV (coefficient of variation) percentages describing the dispersion of results around the mean varied from 1.5% (*g03*) to 49.5% (g77). Majority of the CVs varied between 25% and 40% as the CVs of only four genes (*g03, g05, g04* and *g48)* fell below 10% (Table 5).

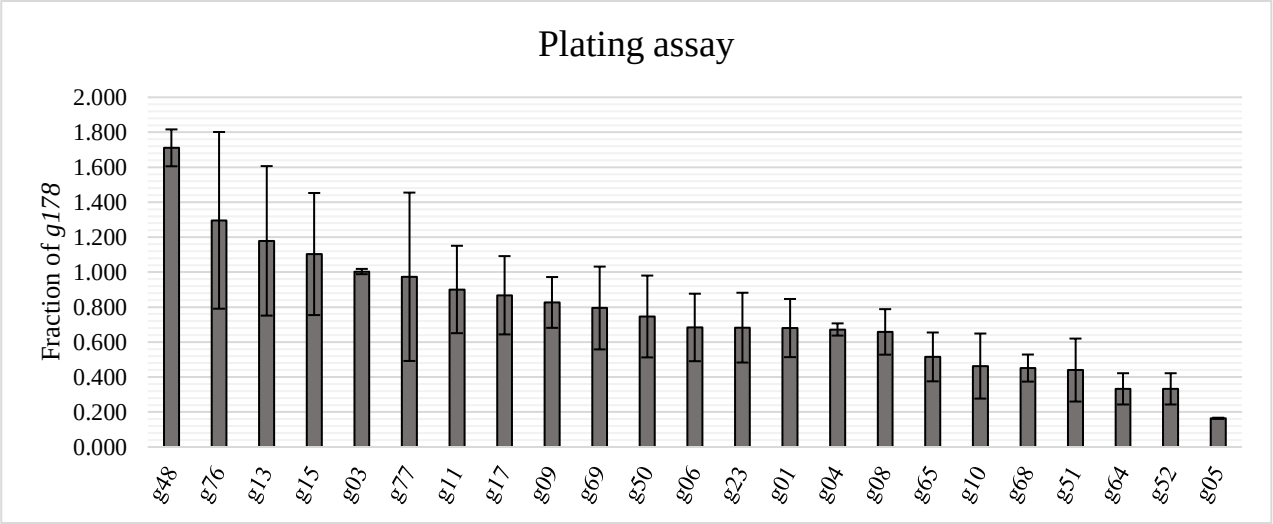

Figure 5. Plating assay results of fHy-Eco03 HPUF genes. Results are shown as average fractions against a non-toxic control gene *g178* (CFU 50 µl) in that set of transformations. The bars stand for mean ± SD between triplicate plating's of two biological replicates.

NGS assay

The total read coverages per gene were obtained as a sum of four sequence fragments as described earlier (Figures 2A and 2B). The amounts of sequence fragments from the transformant derived plasmids were equally represented within genes, but in the ligation mixture, sequence coverages varied consistently up to a thousand-fold between complementary fragments (Table S5). The same trend was seen also in the pre-experiment results (Table S3). Throughout genes the amounts of atg containing forward and stop codon containing reverse sequences were similar whereas only a few and in some cases none of the atg containing reverse fragments were obtained. The average sequence read amounts and proportions from the replicate ligation mixtures and transformant plasmid samples are presented in Table 4. In the ligation mixtures read coverages varied in average from ~150 (0.9%) to ~2000 (12%) and in the transformant samples from ~3000 (1%) to ~2×10^4 (5%) reads per gene, with the exception of *g05* (37 reads, 0.01%) and *g48* (5×10^4 reads, 15%).

The ratios of relative reads presented in Table 5, varied from the lowest of 0.003 of gene *g05* to 5.9 of gene *g51*. The relative amount of gene *g05* reads decreased by 300% from the ligation mixture. The relative amounts of nine other putatively toxic genes decreased between 30% (~0.7, *g17*) and 80% (~0.2, *g64*). The variation between the replicate screenings was minor, as the standard deviations of only three genes were above 0.300 (*g03*, *g06* and *g10*) with majority falling below 0.100. The dispersion of results (coefficient of variation) from the average ratio was under 10% with a vast majority of the genes, as only two genes had CVs over 20% (*g05* 33%, *g10* 28%) (Table 5),

from which the result of gene *g05* is explained by the weaker applicability of CV when means and SDs are low.

Table 4. Sequence read coverages and the corresponding relative amounts (%) of ligation products in the ligation mixtures and transformant plasmid samples. The results are averages of two pairs of replicate transformations and two biological replications of ligation mixtures.

Gene	Ligation mix	Relative amount %	Transformant sample	Relative amount %
g77	1894	11.5	17216	4.8
g76	578	3.5	15974	4.4
g69	1355	8.2	20288	5.6
g68	581	3.5	3833	1.1
g65	894	5.4	19470	5.4
g64	1310	7.9	7480	2.1
g52	1118	6.8	13739	3.8
g51	151	0.9	19382	5.4
g50	442	2.7	13642	3.8
g48	564	3.4	55506	15.4
g23	2003	12.1	11847	3.3
g17	1290	7.8	19801	5.5
g15	1432	8.7	18776	5.2
g13	475	2.9	20594	5.7
g11	177	1.1	10718	3.0
g10	180	1.1	5867	1.6
g09	465	2.8	15512	4.3
g08	212	1.3	13885	3.8
g06	243	1.5	12825	3.6
g05	557	3.4	37	0.01
g04	184	1.1	12748	3.5
g03	209	1.3	19221	5.3
g01	217	1.3	12714	3.5

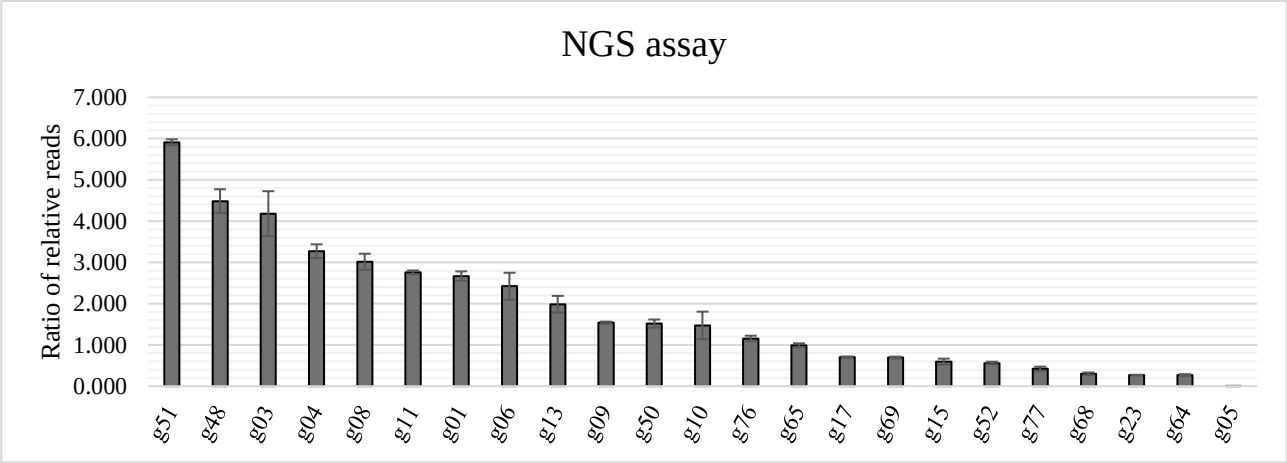

Figure 6. NGS assay results of fHy-Eco03 HPUF genes. The bars stand for mean ± SD between two pairs of replicate transformations from two biological replications of ligation mixtures.

NGS assay vs. plating assay

The NGS and plating assay results presented in Table 5 were mostly in-line with each other, as the distribution of genes along the grade of toxicity was very similar. Six genes from the ten most putatively toxic ones were same in both assays and from these gene *g05* was presented as the most toxic. However, in the NGS assay the difference between gene *g05* and other genes was substantial as the ratio of the second most putatively toxic gene (*g64*) was almost a 100-fold higher. In the plating assay, six of the most putatively toxic genes were in a relatively close range varying from 16% (*g05*) to 40% (*g68*) of the non- toxic control gene. The plating assay results were more dispersed between replicates as can be seen from Table 5 and in detail in Table S4, even the non-toxic control gene showed inconsistent results from batch to batch. In the NGS assay, variation between results was considerably lower not only within but also between the biological replications (Table 5 and Table S5). The CV percentages are, despite a few exceptions, below 10% whereas in the plating assay majority of the results are dispersed from 25% to over 30% around the mean. Based on the results, genes *g05*, *g23*, *g51*, *g52*, *g64* and *g68* showing highest putative toxicities in both assays were chosen for further screening (Table 5).

Table 5. NGS and plating assay results with standard deviations (SD) and coefficient of variation percentages (CV%). Plating assay results are average fractions from triplicate plating's (CFU/50 µl) of two biological replicates against a non-toxic control gene *g178* in that set of transformations. The NGS assay results are averages between two biological replicates of pooled ligation mixtures, from each mixture two replicate transformations were performed. The genes chosen for further screening are highlighted in red.

Plating assay				NGS assay			
Gene	Fraction of *g178*	SD	CV%	Gene	Ratio of relative reads	SD	CV%
g48	1.711	0.106	6.2	*g51*	5.910	0.072	1.2
g76	1.296	0.505	39.0	*g48*	4.484	0.288	6.4
g13	1.179	0.428	36.3	*g03*	4.179	0.546	13.1
g15	1.104	0.349	31.6	*g04*	3.271	0.167	5.1
g03	1.003	0.015	1.5	*g08*	3.013	0.196	6.5
g77	0.974	0.481	49.5	*g11*	2.760	0.043	1.6
g11	0.901	0.250	27.8	*g01*	2.668	0.114	4.3
g17	0.868	0.224	25.8	*g06*	2.421	0.328	13.5
g09	0.827	0.145	17.6	*g13*	1.986	0.202	10.2
g69	0.795	0.237	29.8	*g09*	1.537	0.027	1.7
g50	0.746	0.234	31.4	*g50*	1.516	0.099	6.5
g06	0.684	0.193	28.2	*g10*	1.471	0.334	22.7
g23	0.683	0.200	29.3	*g76*	1.155	0.068	5.9
g01	0.680	0.166	24.4	*g65*	0.989	0.049	4.9
g04	0.672	0.035	5.2	*g17*	0.705	0.014	2.0
g08	0.658	0.130	19.8	*g69*	0.693	0.023	3.3
g65	0.515	0.140	27.1	*g15*	0.593	0.074	12.5
g10	0.463	0.186	40.2	***g52***	**0.559**	**0.033**	**5.9**
g68	**0.451**	**0.077**	**17.1**	*g77*	0.422	0.051	12.1
g51	**0.440**	**0.180**	**40.9**	***g68***	**0.298**	**0.034**	**11.4**
g64	**0.332**	**0.089**	**26.9**	***g23***	**0.271**	**0.005**	**2.0**
g52	**0.332**	**0.089**	**26.9**	***g64***	**0.267**	**0.024**	**9.1**
g05	**0.164**	**0.003**	**2.1**	***g05***	**0.003**	**0.001**	**33.4**

Confirmation of toxicity

Six genes (*g05*, *g23*, *g51*, *g52*, *g64* and *g68*) producing the lowest sequence read ratios and relative fractions were chosen for the growth curve analysis to confirm toxicities. From these genes, *g05*, *g64*, *g68*, and *g52* were among the putative hits in both assays (Table 5). Figures 8 and 9 show the absorbances (OD_{600}) measured at an hourly rate during a 20-hour incubation. During arabinose induced expression, the growth curve of gene *g05* containing cells resembles that of toxic control

gene (ϕR1-RT *g137*), with an OD_{600} below 0.1 – 0.2 throughout the incubation. The growth curves of genes *g64* and *g51* cells are similar with non-toxic (ϕR1-RT *g150*) and vector (pBAD30) controls, showing a steady exponential increase in absorbance (<0.8) over time. Growth curves of cells containing genes *g52* and *g23* show slight decrease in growth rate settling between the final OD_{600} of 0.6 – 0.7. The growth curve of gene *g68* containing cells has a more linear form, barely reaching the absorbance of 0.6 at the final timepoint (Figure 7). During glucose inhibited protein expression, absorbances rise exponentially over the 20-hour period, with no remarkable signs of decline or differentiation between genes, as was expected. Only a very slight aberration in the curves of genes *g68*, *g23* and *g05* containing cells is seen, probably the result of leaked expression from the promoter (Figure 8).

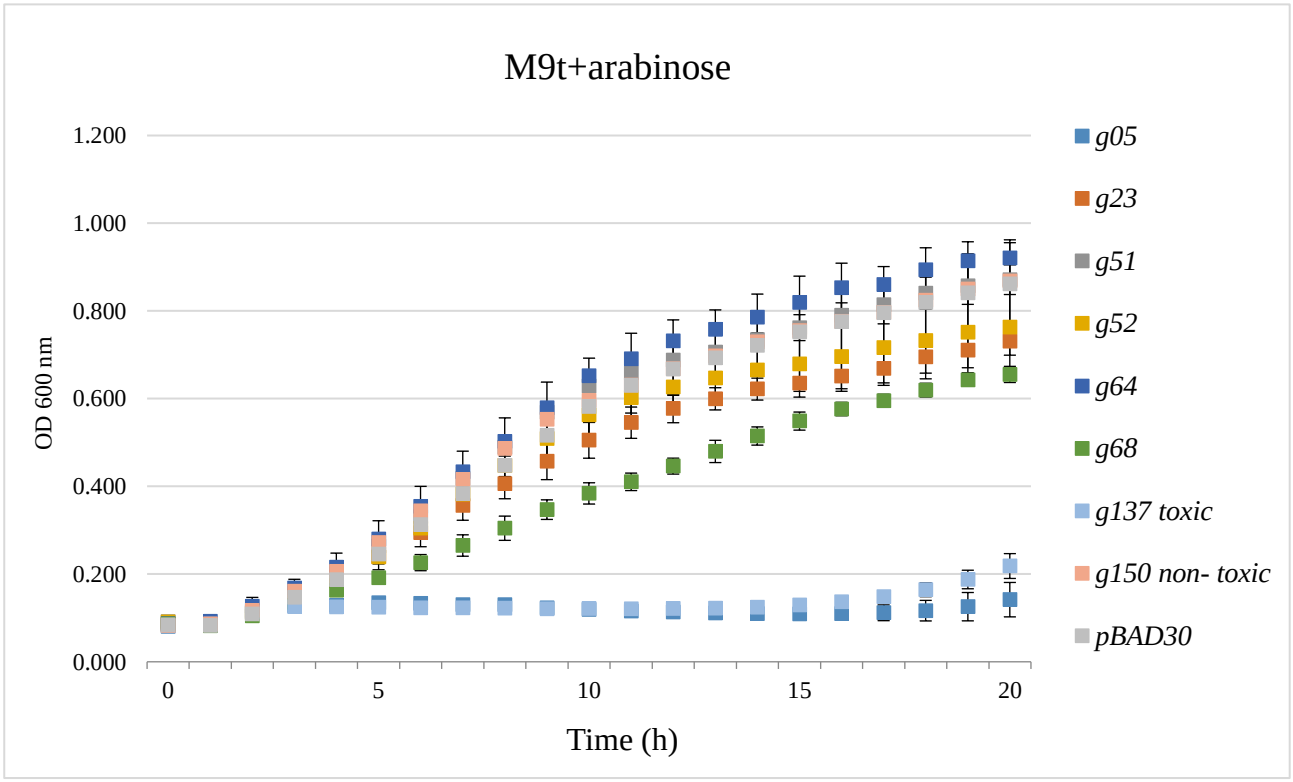

Figure 7. Growth curves of *E. coli* DH5α cells grown in M9t supplemented with arabinose. The curves are average ODs (600 nm) from triplicate wells of three replicates.

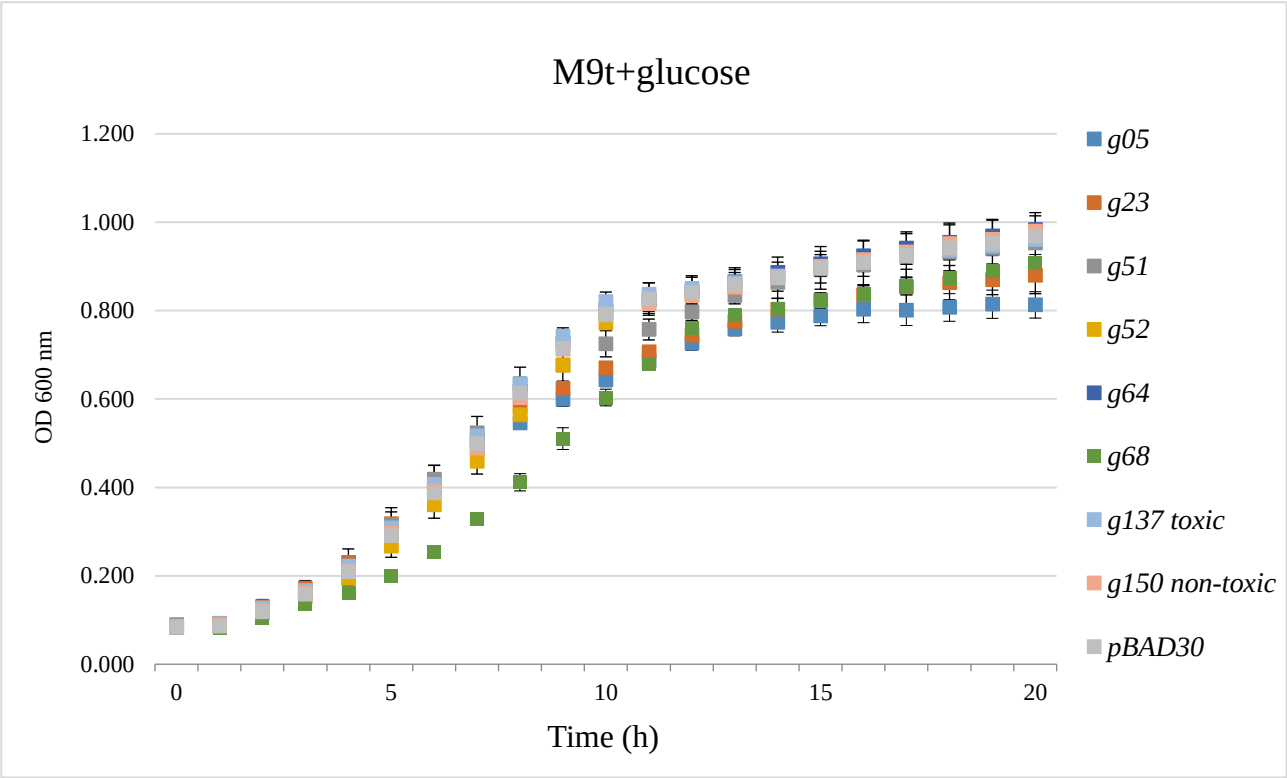

Figure 8. Growth curves of *E. coli* DH5α cells grown in M9t supplemented with glucose. The curves are average ODs (600 nm) from triplicate wells of three biological replicates.

Protein and sequence analysis of g05

The protein sequence of Gp05 was aligned against protein databases with a BLASTx query. The sequence had similarities against five *Salmonella* phage proteins annotated as hypotheticals with query coverages varying from 91% to 98%. The highest sequence similarity of 83% was against a hypothetical protein of *Salmonella* phage strain SE4 (MK770413). To identify the possible mechanisms of toxicity, functional domains and secondary structures, the protein structure of gene *g05* was modelled using Phyre2 software. From the 81 amino acids of Gp05, 39 residues were modelled with a 73.60% confidence as helicase RecG (PDB 1GM5) (Figure 9) of *Thermotoga maritima* (Singleton et al., 2001). The sequence coverage of 31%, however, is relatively low for a reliable prediction.

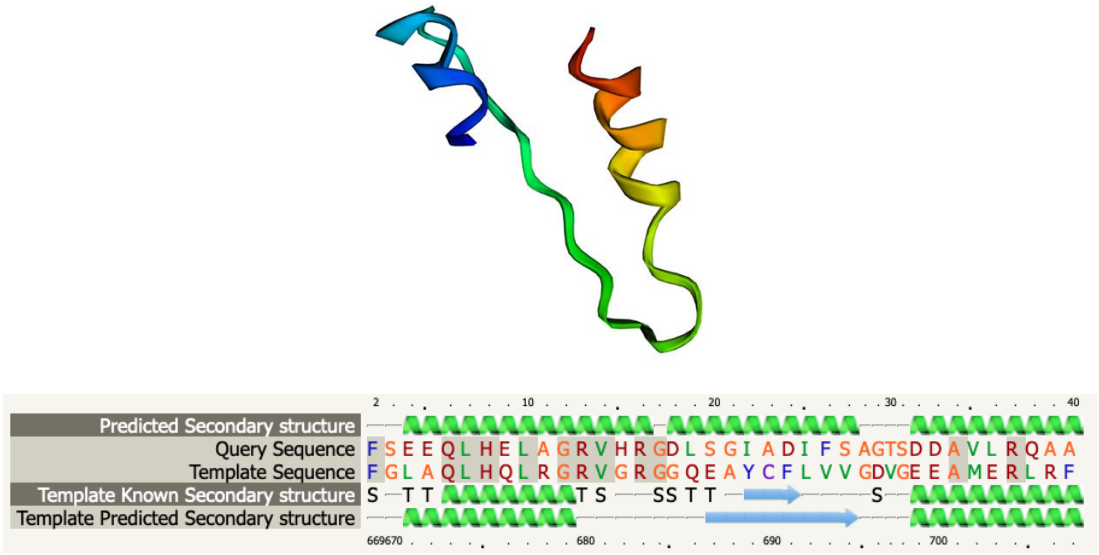

Figure 9. Predicted protein structure and protein sequence alignment of Gp05, modelled with Phyre2 software.

DISCUSSION

The ever-increasing abundance of antibiotic resistant bacteria creates a constant demand for new drugs to replace those that have lost the ability to combat infectious bacteria. Drug development nonetheless is always one step behind, as the discovery and approval of new antibiotics cannot keep up with the rapid pace of bacterial evolution (Genilloud, 2017). The long-term strategy in drug development has for long relied on the existing antibiotics as a reference for new molecules. This, however, is not an inexhaustible pool and alternative sources are needed. The hypothetical proteins produced during phage infection cycle could be utilized in the discovery of new molecules used for antibacterial purposes in the future (Schmelcher et al., 2012; Van Den Bossche et al., 2014; Roach & Donovan, 2015). However, the current methods used for the screening of such proteins are both time consuming and tedious, making this an unfavorable and high-cost option to capitalize, creating the need for a new more efficient and reliable method.

In this study, a new NGS based assay for the screening of phage derived bacteriotoxic proteins was developed and tested. Next Generation Sequencing is used to overcome the challenges with reliability and reproducibility, but also to reduce the amount of time and resources required in formerly used assays. One of these assays (Moharanj et al., 2019) was chosen as a reference study to compare the performance against. In this so-called plating assay, the workload was optimized by first

selecting true hypotheticals through proteomics and annotation, and three toxic genes were found by screening 129 HPUFs of phage ϕR1-RT. Reliable results were obtained by performing three replicate transformations from three biological replications of each gene. While this does minimize the issue with reliability; it drastically increases the amount of time and labor spent.

Preliminary experiment

In the NGS assay, the obstacles discussed earlier are avoided by firstly performing transformations as a pooled sample and secondly by obtaining the amount of correctly ligated plasmids thru bioinformatics, thus excluding any false interpretations caused by variations between replicates and incorrectly ligated or digested plasmids. As a proof of principle, a preliminary experiment of the NGS assay was performed with five previously identified and confirmed, toxic and non-toxic proteins of phages ϕR1-RT (Moharanj et al., 2019), fHe-Kpn01 (Spruit et al., 2020) and T4 (Ruckman et al., 1994). The amounts of sequence reads were expected to correlate with the toxicities of the genes, as an outcome of decrease or increase in growth of host bacteria upon protein expression. This however was not verified as only *g22* and *g10* appeared to have an inhibitory impact on bacterial growth (Figure 5). Toxic gene *regB* fragments were not detected from any of the samples, probably due to human error, either in primer selection or preparation of the ligation mixture. The plating assay gave results in-line with the toxicities of the genes, whereas no reliable conclusions about the toxicities of the genes could be drawn from the NGS assay results.

Two possible factors explaining the inconsistent results were hypothesized. The first one is the possible sufficient repression of *lac* promoter located upstream to the multiple cloning site of plasmid pU11L4. Plasmid pU11L4, used as a cloning vector in this and also the comparative study done by Mohanraj et al., (2019) is a pUC19 based vector. PUC based vectors are one of the most commonly used plasmids in cloning and protein expression studies. Incomplete repression and transcriptional read through makes pUC based high copy number plasmids good candidates for screening assays such as this identifying toxic genes (Mohanraj et al., 2019). Protein expression in these vectors is controlled by allosteric regulation of the *lac* promoter, repressed with high levels of glucose and induced in the presence of lactose or IPTG (Isopropyl-β-d-1-thiogalactopyranoside). However, in the absence of an inducer, protein expression often occurs as a result of high basal expression (Rosano & Ceccarelli 2014). Glucose is often used as a supplement in growth media to enhance the recovery of cells after electroporation (van der Rest et al., 1999). Growth media, containing a 0.2% glucose supplement was used in the incubation of cells after electroporation but also to further enumerate the transformant colonies before plasmid isolations in this study. Previous

research indicates that a concentration of 0.4% of glucose is enough to inhibit expression, while with lower concentrations from 0.05%-0.1% overproduction of proteins can occur. Glucose is sufficient to repress expression during log phase, after which it usually has been consumed and basal level expression is restored. (De Bellis & Schwartz, 1990). Glucose supplementation used in this study may have been enough to repress expression to the level where only those proteins (*g10, g22*) toxic at very low amounts inhibited bacterial growth (Figure 4).

The second challenge discovered in the analysis of results was the use of relative ratio as an indicator of toxicity since the number of sequence reads obtained from the ligation mixture varied considerably (Table 3). This was seen especially with toxic gene *g137* that was highly overrepresented in the results due to the very low proportional amount of sequence reads obtained from the ligation mixture sample. The disproportion of successfully cloned plasmids in the pre-electroporation sample may thus affect interpretations about the level of toxicities. However, it has to be noted, that the inconsistent results are likely the cause of repression related issues and without this, the variable sequence read coverages in the ligation mixture would not have affected the results at this level.

Sequence analysis

During cloning, a variety of combinations between inserts and plasmids can occur, so the ligation sample contains not only successfully cloned plasmids but also a variation of DNA constructs such as, double ligated inserts, undigested plasmids, non-ligated inserts and linear plasmids (Figure 2A). From these, undigested plasmids can easily be interpreted as false positives in the plating assay. Because of this, plasmids were searched by aligning cloning site fragments from both strands, containing start and stop codons of the genes, thus showing that the samples contain intact and correctly ligated plasmids (Figure 2A, 2B). The amount of sequence reads of complementary strands in the transformant derived plasmids were similar, but for some reason the corresponding reads in the ligation mixture were highly mismatched as can be seen from Tables S3 and S5. It appears as though the ligation mixture samples would consist mostly of partially cloned linear plasmids and only a few and in some cases none of the double stranded plasmids. However, since the read amounts of these fragments in the transformant sample sequences are equal and fairly high, it’s very unlikely that the sample would not contain intact double stranded plasmids. The reasons behind these results remain unclear.

Screening of fHy-Eco03

The performance of the NGS assay was further tested by screening 23 previously identified hypothetical ORFs of phage fHy-Eco03. The possible glucose associated issues seen in the pre-experiment may have occurred also in this screening. Despite this, results between the NGS- and plating assay were surprisingly in-line with each other, with many similarities in the composition of genes presenting toxic properties (Figure 5 Figure 6). Six HPUFs were the same among the ten most putatively toxic genes in both assays, and a similar trend could be seen in the genes considered as non-toxic (Table 5). It might be that the higher abundance of transformant bacteria were sufficient enough to reduce the effects of glucose more efficiently than in the pre-experiment, therefore enabling basal expression. Gene *g05* identified as toxic inhibited growth most effectively in both assays, but in the NGS assay the decline was substantial (Table 5). In addition to this, variations within but also between the replicate screenings were considerably low indicating high reliability and reproducibility. Despite the similarity of results, the challenges of the plating assay, reviewed earlier in detail were present also in this study. The plating assay results were more inconsistent varying between replicate transformations at highest over two-fold and similar variation was seen also with the non-toxic control from batch to batch (Table S4). The results were most probably affected by variations originating from various sources and steps, from cloning all the way to the plating of cells.

Based on the screening assay results, the toxic properties of genes *g05*, *g23*, *g51*, *g52*, *g64* and *g68* were further analyzed with inducible expression plasmids. In this study the primary screening and confirmation was performed in the host species of phage fHy-Eco03, thus toxicities were expected to apply in both the preliminary and confirmatory screening. However, if the host cells used are not the natural targets of the protein encoding bacteriophage, confirmation of toxicity in the target host is advised. Gp05 of fHy-Eco03, confirmed as a toxic, was further analyzed to identify close analogies in databases and the predicted molecular structure. The sequence had over 80% identity against hypothetical proteins of five *Salmonella* phages, but no other significant similarities were discovered. *E. coli* and *Salmonella* species are indeed closely related (Hu et al., 2010) hence it is possible that the structure and function of Gp05 is similar with the hypothetical proteins obtained in the query. The protein structure of Gp05 was predicted to contain helicase RecG domain, however the required 90% confidence for a reliable prediction was not met and the sequence coverage was just above the recommended limit of 30% (Kelley et al., 2015). RecG is a multifunctional bacterial protein associated with DNA repair, unwinding of R-loops, adaptation to CRISPR-Cas, replication initiation and repair of replication forks (Dudas, & Kreuzer, 2001; Lloyd & Rudolph, 2016). Even though these results are not enough for a reliable prediction about the structure or function of Gp05, some

speculation could be made about the possible inhibition of RecG targets or RecG itself, since many of the cellular purposes of this protein are essential for successful DNA replication. Furthermore, a detailed molecular analysis should be conducted in order to define the exact function and targets of Gp05 and whether these properties appear toxic across bacterial species. This requires further protein-protein interaction studies with affinity tagged proteins, although this is fairly challenging since the production of toxic proteins is often lethal to the bacterial host (Spruit et al., 2020).

Possible challenges

Although the NGS assay is designed to minimize false interpretations caused by variations in sample composition and lab work, a few possible issues were hypothesized. One concern was that the over- or under-representation of recombinant plasmids in the ligation mixture could affect the amounts in the transformed cells. This however does not seem to be the case since the proportions of sequences between these two samples do not correlate, although it cannot be confirmed that the recombinant plasmids are represented in the same equivalent proportions during transformation. The disproportionate representation of ligation products was more of a problem in the analysis of the results as was seen in the preliminary experiment. The second significant factor in the screening assay is the transformation efficiency of the host cells (Tu et al., 2005). In order to obtain sufficient coverages of genes, the number of pooled colonies should be adequate. In this study, approximately 10^4 (pre-experiment) to 2×10^4 (fHy-Eco03 screening) colonies per transformation were pooled resulting in approximately 100x coverage per gene. The third issue was the possible effect of plasmid size on transformation efficiency. The causation of plasmid size and transformation efficiency has been studied previously, and some declines in efficiencies were observed with increased plasmid sizes (Ohse et al., 1995). However, the variation between sizes were far greater than in this study where the biggest gene was ~1700 bp (*g150*), not having a significant impact on the recombinant plasmid size. Declines in transformation efficiencies were previously observed with plasmids over 10kb in size (Ohse et al., 1995). Based on the results of this study, plasmid size or plasmid amounts, and sequence read coverages did not correlate, though it is noteworthy that if some differences occurred, they were most probably diminished during the incubation of transformant cells prior to plasmid isolations.

Conclusions and future prospects

Although the results in this study were most likely affected by repression related issues, the performance of the NGS assay proved to outdo the plating assay in efficiency, accuracy and reliability, especially based on the screening results of fHy-Eco03. Sequencing and bioinformatics enable the identification of toxic genes with high accuracy and negligible deviations between replications. Compared to previous assays, the sample size of transformations is diminished from tens if not hundreds to a minimum of two samples, regardless of the number of genes, thus saving both time and resources. The individual screening of genes in the previous assays may take several weeks to even months to finish. After successful cloning, the manual work of the NGS assay can be performed in a matter of few days. The sequence data can be processed and analyzed with minor effort through a pipeline provided in the supplementary material S6. The genes presenting antibacterial properties in this study were most likely highly toxic, and it is uncertain whether the use of glucose supplementation inhibited identification of less toxic genes. In the future, repression associated issues can be avoided by selecting appropriate growth media and by including both toxic and non-toxic control genes in the screening. The results can thus be obtained in addition to the relative ratio, by also comparing the results against known toxic and non-toxic controls. The controls should optimally include genes with different levels of toxicities to ensure that not only those genes with high antibacterial properties are identified, and also to control that protein expression occurs. The NGS assay should be further validated by optimization and re-testing with toxic and non-toxic genes but also by experimentation with various bacteriophages and host species. Proteins with conserved cellular targets across species or genera, could be identified by screening the genes in both gram- negative and positive bacteria.

REFERENCES

Ackermann, HW (2011) Bacteriophage taxonomy. *Microbiol. Aust*, *32*(2), 90-94. doi: 10.1071/MA11090.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J. *Mol. Biol*. 215:403-410. doi: 10.1016/S0022-2836(05)80360-2

Bush K, (2010) Alarming β-lactamase-mediated resistance in multidrug-resistant Enterobacteriaceae. *Current Opinion in Microbiology*, *13*(5), 558-564. doi: 10.1016/j.mib.2010.09.006.

Campos-Guillén J, Fernández F, Pastrana X, Loske AM (2012) Relationship between plasmid size and shock wave-mediated bacterial transformation. *Ultrasound in Medicine & Biology*, *38*(6), 1078-1084. doi: 10.1016/j.ultrasmedbio.2012.02.018.

Cooper MA & Shlaes, D (2011) Fix the antibiotics pipeline. *Nature*, *472*(7341), 32-32. doi: 10.1038/472032a.

Coque TM, Baquero F, & Canton R (2008) Increasing prevalence of ESBL-producing Enterobacteriaceae in Europe. *Eurosurveillance*, *13*(47), 19044. Doi: 10.2807/ese.13.47.19044-en

De Bellis D & Schwartz I (1990) Regulated expression of foreign genes fused to lac: control by glucose levels in growth medium. *Nucleic Acids Research*, *18*(5), 1311. doi: 10.1093/nar/18.5.1311.

Dower WJ, Miller JF, Ragsdale CW (1988) High efficiency transformation of E. coli by high voltage electroporation. *Nucleic Acids Research*, *16*(13), 6127-6145. doi: 10.1093/nar/16.13.6127.

Dudas KC & Kreuzer KN (2001) UvsW protein regulates bacteriophage T4 origin-dependent replication by unwinding R-loops. *Molecular and Cellular Biology*, *21*(8), 2706-2715. doi: 10.1128/MCB.21.8.2706-2715.2001.

Endersen L, O'Mahony J, Hill C, Ross, RP, Mc.Auliffe O, Coffey A (2014) Phage therapy in the food industry. *Annual Review of Food Science and Technology*, *5*, 327-349. doi: 10.1146/annurev-food-030713-092415.

Enright MC, Robinson DA, Randle G, Feil EJ, Grundmann H, Spratt BG (2002) The evolutionary history of methicillin-resistant Staphylococcus aureus (MRSA). *Proceedings of The National Academy Of Sciences*, *99*(11), 7687-7692. doi: 10.1073/pnas.122108599.

Fernandes P, Martens E (2017) Antibiotics in late clinical development. *Biochemical Pharmacology*, *133*, 152-163. doi: 10.1016/j.bcp.2016.09.025.

Genilloud O (2017) Actinomycetes: still a source of novel antibiotics. *Natural Product Reports*, *34*(10), 1203-1232. doi: 10.1039/C7NP00026J.

Hendrix RW (2002) Bacteriophages: evolution of the majority. *Theoretical Population Biology*, *61*(4), 471-480. doi: 10.1006/tpbi.2002.1590.

Hu B, Perepelov AV, Liu, B, Shevelev SD, Guo D, Senchenkova, SYN et al. (2010). Structural and genetic evidence for the close relationship between Escherichia coli O71 and Salmonella enterica O28 O-antigens. *FEMS Immunology & Medical Microbiology*, 59(2), 161-169. doi 10.1111/j.1574-695X.2010.00676.x

Inoue H, Nojima H, Okayama H (1990) High efficiency transformation of Escherichia coli with plasmids. *Gene*, *96*(1), 23-28. doi: 10.1016/03781119(90)90336-P.

Jorge P, Pérez-Pérez M, Rodríguez GP, Pereira MO, Lourenço A (2017) A network perspective on antimicrobial peptide combination therapies: the potential of colistin, polymyxin B and nisin. *International Journal of Antimicrobial Agents*, *49*(6), 668-676. doi: 10.1016/j.ijantimicag.2017.02.012.

Karthik K, Muneeswaran NS, Manjunathachar HV, Gopi M, Elamurugan A, Kalaiyarasu S (2014) Bacteriophages: effective alternative to antibiotics. *Adv. Anim. Vet. Sci*, *2*(3S), 1-7. doi: 10.14737/journal.aavs/2014/2.3s.1.7.

Kelley LA, Mezulis, S, Yates CM, Wass MN, Sternberg MJ (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* 10, 845-858 doi: 10.1038/nprot.2015.053.

Lamberg A, Nieminen S, Qiao M, Savilahti H (2002) Efficient insertion mutagenesis strategy for 552 bacterial genomes involving electroporation of in vitro-assembled DNA transposition complexes of 553 bacteriophage mu. *Appl Environ Microbiol*, 68, 705-712.Doi: 10.1128/AEM.68.2.705-712.2002

Laxminarayan R, Duse A, Wattal C, Zaidi AK, Wertheim HF, Sumpradit N, Greko C (2013) Antibiotic resistance—the need for global solutions. *The Lancet Infectious Diseases*, *13*(12), 1057-1098. doi: 10.1016/S1473-3099 (13) 70318-9.

Liu J, Dehbi M, Moeck G, Arhin F, Bauda P, Bergeron D, McCarty J (2004) Antimicrobial drug discovery through bacteriophage genomics. *Nature Biotechnology, 22*(2), 185-191. doi: 10.1038/nbt932.

Lloyd RG, Rudolph CJ (2016) 25 years on and no end in sight: a perspective on the role of RecG protein. *Current Genetics*, *62*(4), 827-840. doi: 10.1007/s00294-016-0589-z.

Mohanraj U, Wan X, Spruit CM, Skurnik M, Pajunen MI (2019) A Toxicity Screening Approach to Identify Bacteriophage-Encoded Anti-Microbial Proteins. *Viruses*, *11*(11), 1057 doi: 10.3390/v11111057.

Munita JM, Arias CA (2016) Mechanisms of antibiotic resistance. *Virulence Mechanisms of Bacterial Pathogens*, 481-511. doi: 10.1128/9781555819286.ch17.

Mushegian AR (2020) Are There 1031 Virus Particles on Earth, or More, or Fewer? *Journal of Bacteriology*, 202(9). Doi: 10.1073/pnas

Nielsen TB, Brass EP, Gilbert DN, Bartlett JG, Spellberg B (2019) Sustainable discovery and development of antibiotics—is a nonprofit approach the future?. *The New England Journal of Medicine, 381*(6), 503. doi: 10.1056/NEJMp1905589.

Ohse M, Takahashi K, Kadowaki Y, Kusaoke H (1995) Effects of plasmid DNA sizes and several other factors on transformation of Bacillus subtilis ISW1214 with plasmid DNA by electroporation. *Bioscience, Biotechnology, and Biochemistry*, *59*(8), 1433-1437. doi: 10.1271/bbb.59.1433.

Parisien A, Allain B, Zhang J, Mandeville R, Lan CQ (2008) Novel alternatives to antibiotics: bacteriophages, bacterial cell wall hydrolases, and antimicrobial peptides. *Journal Applied Microbiology*, *104*(1), 1-13. doi: 10.1111/j.1365-2672.2007.03498.x

Petty NK, Zakour NLB, Stanton-Cook M, Skippington E, Totsika M, Forde BM, Rogers BA (2014) Global dissemination of a multidrug resistant Escherichia coli clone. *Proceedings of the National Academy of Sciences, 111*(15), 5694-5699. doi: 10.1073/pnas.1322678111.

Rice LB (2006) Antimicrobial resistance in gram-positive bacteria. *American Journal of Infection Control*, *34*(5), 11-19. doi: 10.1016/j.ajic.2006.05.220.

Roach DR, Donovan DM (2015) Antimicrobial bacteriophage-derived proteins and therapeutic applications. *Bacteriophage*, *5*(3), e1062590. doi: 10.1080/21597081.2015.1062590.

Rogers BA, Sidjabat HE, Paterson DL (2011) Escherichia coli O25b-ST131: a pandemic, multiresistant, community-associated strain. *Journal of Antimicrobial Chemotherapy*, *66*(1), 1-14. doi: 10.1093/jac/dkq415.

Rosano GL & Ceccarelli EA (2014) Recombinant protein expression in Escherichia coli: advances and challenges. *Frontiers in Microbiology*, 5, 172. doi: 10.3389/fmicb.2014.00172

Ruckman J, Ringquist S, Brody E, Gold L (1994) The bacteriophage T4 regB ribonuclease. Stimulation of the purified enzyme by ribosomal protein S1. *Journal of Biological Chemistry*, *269*(43), 26655-26662.

Rupp ME, Fey PD (2003) Extended spectrum β-lactamase (ESBL)-producing Enterobacteriaceae. *Drugs*, *63*(4), 353-365. doi: 10.2165/00003495-200363040-00002.

Salmond GP, Fineran PC (2015) A century of the phage: past, present and future. *Nature Reviews Microbiology*, *13*(12), 777-786. doi: 10.1038/nrmicro3564.

Schmelcher M, Donovan DM, Loessner MJ (2012) Bacteriophage endolysins as novel antimicrobials. *Future Microbiology*, *7*(10), 1147-1171. doi: 10.2217/fmb.12.97.

Schmelcher M, Loessner MJ (2016) Bacteriophage endolysins: applications for food safety. *Current Opinion in Biotechnology*, *37*, 76-87. doi: 10.1016/j.copbio.2015.10.005.

Shibayama Y, Dabbs ER (2011) Phage as a source of antibacterial genes: Multiple inhibitory products encoded by Rhodococcus phage YF1. *Bacteriophage*, *1*(4), 195-197. doi: 10.4161/bact.1.4.17746.

Singh S. Godavarthi S, Kumar A, Sen RA (2019) mycobacteriophage genomics approach to identify novel mycobacteriophage proteins with mycobactericidal properties. *Microbiology*, *165*, 722–736. doi: 10.1099/mic.0.000810.

Singleton MR, Scaife S, Wigley DB (2001) Structural analysis of DNA replication fork reversal by RecG. *Cell*, 107(1), 79-89. Doi: 10.1016/S0092-8674(01)00501-3

Tacconelli E, Carrara E, Savoldi A, Harbarth S, Mendelson M, Monnet DL, Ouellette M et al. (2018) Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *The Lancet Infectious Diseases*, *18*(3), 318-327. doi: 10.1016/S1473-3099(17)30753-3.

Tu Z, He G, Li KX, Chen MJ, Chang J, Chen L et al. (2005). An improved system for competent cell preparation and high efficiency plasmid transformation using different Escherichia coli strains. *Electronic Journal of Biotechnology*, 8(1), 113-120.

Van Den Bossche A, Ceyssens PJ, De Smet J, Hendrix H, Bellon H, Leimer N, et al. (2014). Systematic identification of hypothetical bacteriophage proteins targeting key protein complexes of Pseudomonas aeruginosa. *Journal of Proteome Research*, *13*(10), 4446-4456. doi: 10.1021/pr500796n.

Van der Rest ME, Lange C, Molenaar D (1999) A heat shock following electroporation induces highly efficient transformation of Corynebacterium glutamicum with xenogeneic plasmid DNA. *Applied Microbiology and Biotechnology*, 52(4), 541-545. doi: 10.1007/s002530051557

Von Bubnoff A (2008) Next-generation sequencing: the race is on. *Cell*, *132*(5), 721-723. doi: 10.1016/j.cell.2008.02.028.

Wicklund A (2014) Kliinisille ESKAPEE-bakteerikannoille spesifisten bakteriofagien eristäminen ja karakterisointi faagiterapiaa varten. Master's thesis. University of Helsinki.

Young RY (2002) Bacteriophage holins: deadly diversity. *Journal of Molecular Microbiology and Biotechnology*, *4*(1), 21-36.

Supplementary materials

Table S1. Primers used to amplify fHy-Eco03 HPUF genes by PCR

Gene	Size (bp)	Primer sequence (5' – 3')	Restriction site
Gp77E-FW	360	GCAGCGGCCGCATGAAAACAATGTTTGTATACAAG	*NotI*
Gp77E-RV		GGTCCATGGTCACAGGCCTAACACCTC	*NcoI*
Gp76E-FW	240	GCAGCGGCCGCATGAAACATCAGGGCACA	*NotI*
Gp76E-RV		GGTCCATGGTTATCTCCAGCGTGCATG	*NcoI*
Gp69E-FW	423	GCAGCGGCCGCATGGAAGCCACCAAGTCT	*NotI*
Gp69E-RV		GGTCCATGGTTATTTATACACTTCAATGGTACAC	*NcoI*
Gp68E-FW	231	GCAGCGGCCGCATGTATACGTTCAATCAATTCAAA	*NotI*
Gp68E-RV		GGTCCATGGTTAATACCCCTCTGTTAACATC	*NcoI*
Gp65E-FW	450	GCAGCGGCCGCATGCAGAACCTTTGGGATA	*NotI*
Gp65E-RV		GGTCCATGGTCAGCTTGTGTAGCCGG	*NcoI*
Gp64E-FW	549	GCAGCGGCCGCATGTTTGTGCCGATTCCG	*NotI*
Gp64E-RV		GGTCCATGGTCAATTCGCTTCTTCAGCG	*NcoI*
Gp52E-FW	276	GCAGCGGCCGCGTGTCTAATATAGGTTTAGAGTACC	*NotI*
Gp52E-RV		GGTCCATGGTCATTTGCATTCTTCCTCG	*NcoI*
Gp51E-FW	156	GCAGCGGCCGCATGAGAAAGTTTACCCTGGA	*NotI*
Gp51E-RV		GGTCCATGGTCAATCATATTTCTTAGCCAAGT	*NcoI*
Gp50E-FW	225	GCAGCGGCCGCATGATTGAATCTACACGCAC	*NotI*
Gp50E-RV		GGTCCATGGTTAGCGCTTTCCTTTCTTTT	*NcoI*
Gp48E-FW	216	GCAGCGGCCGCTTGACCTCATGCAGTAAAAAG	*NotI*
Gp48E-RV		GGTCCATGGTTAATTGGCCCCTTCACT	*NcoI*
Gp23E-FW	369	GCAGCGGCCGCATGGGTGGCGCAAGAATC	*NotI*
Gp23E-RV		GGTCCATGGCTAATTTTGTACAAGTAAACCAGAG	*NcoI*
Gp17E-FW	291	GCAGCGGCCGCATGGACTTATTCGACTTAGTTGA	*NotI*
Gp17E-RV		GGTCCATGGTTATGCCATTTCTACCAGCA	*NcoI*
Gp15E-FW	507	GCAGCGGCCGCATGAACATCATAGTTGAGGGA	*NotI*
Gp15E-RV		GGTCCATGGTCATTTAGTCCACTCCTTATCG	*NcoI*
Gp13E-FW	237	GCAGCGGCCGCATGACTGTAGTTGCACCT	*NotI*
Gp13E-RV		GGTCCATGGTCATTTTACCTTAAATTTAAGCCA	*NcoI*

Gp11E-FW	171	GCAGCGGCCGCATGCATACTCAATTAATCATCTGG	*NotI*
Gp11E-RV		GGTCCATGGTCATTCTCCTGAGTTAAATTGTG	*NcoI*
Gp10E-FW	177	GCAGCGGCCGCATGATGAAAGAATTGACATTGAC	*NotI*
Gp10E-RV		GGTCCATGGCTATTGTAACTTTGTACTAATTTGA	*NcoI*
Gp09E-FW	228	GCAGCGGCCGCATGGTAACATTCACAACTATCC	*NotI*
Gp09E-RV		GGTCCATGGTTAAACCTCAAATTTTGATTCGTC	*NcoI*
Gp08E-FW	201	GCAGCGGCCGCATGCCATTAATCAAAGTTACAG	*NotI*
Gp08E-RV		GGTCCATGGTCAAGAAATAATTACCTCAGTAAC	*NcoI*
Gp06E-FW	120	GCAGCGGCCGCATGAAAAAGCATGTTGTTGAG	*NotI*
Gp06E-RV		GGTCCATGGTCAGAAAACATAAGACAACACC	*NcoI*
Gp05E-FW	246	GCAGCGGCCGCATGTTTTCTGAGGAGCAACT	*NotI*
Gp05E-RV		GGTCCATGGTTATTTGCTCCTTACTCCAAGT	*NcoI*
Gp04E-FW	180	GCAGCGGCCGCATGTTTAAATTCTTACAACGTAACC	*NotI*
Gp04E-RV		GGTCCATGGTCACAAGTAATCTCCTTCAAC	*NcoI*
Gp03E-FW	186	GCAGCGGCCGCGTGATTGGCTTAATCCTGG	*NotI*
Gp03E-RV		GGTCCATGGTTAGTCTTGTTTCAACTCATCG	*NcoI*
Gp01E-FW	198	GCAGCGGCCGCATGAGCTATACTGACCAACA	*NotI*
Gp01E-RV		GGTCCATGGTTACAGTTGACCTCTCAGG	*NcoI*

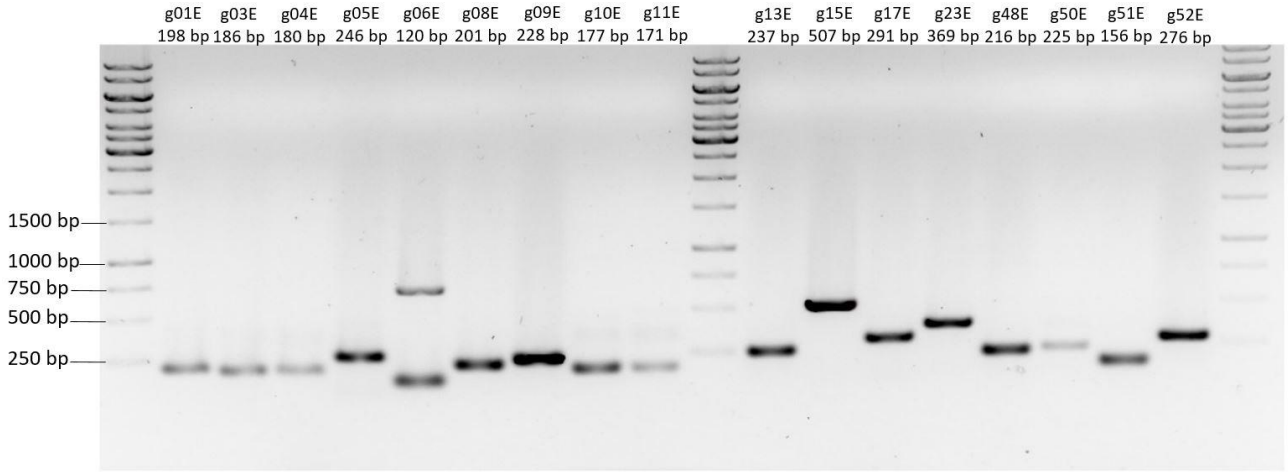

Figure S1a. Gel image of amplified phage fHy-Eco03 genes with a 1kb size marker

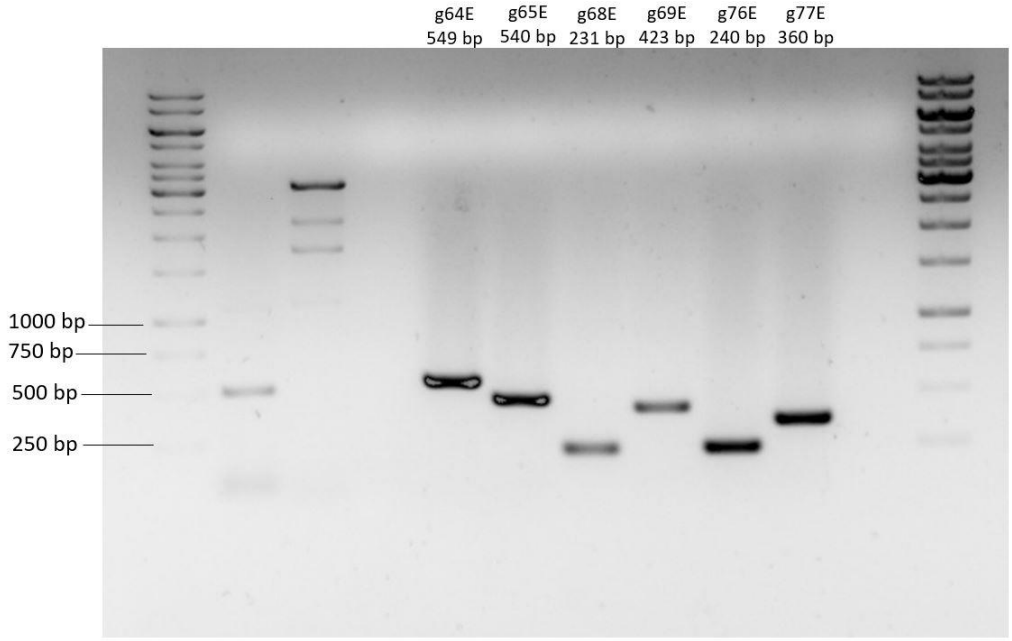

Figure S1b. Gel image of amplified phage fHy-Eco03 genes with a 1kb size marker

Table S2. Plating assay results of control genes.

		CFU1	CFU2	CFU3	Average	SD	CFU1	CFU2	CFU3	Average	SD
Non-toxic	*g178*	760	624	788	724	72	451	391	447	577	157
	g119	350	401	352	368	24	742	797	794	573	206
	g121	366	335	320	340	19	375	378	312	348	26
	g246	247	278	246	257	15	350	376	332	305	51
	g150	199	200	189	196	5	356	312	376	272	78
Toxic	*g137*	85	103	125	104	16	114	122	104	109	13
	g38	65	57.5	93	72	15	78	54	77	71	13
	g22	16	17	16	16	0	12	10	19	15	3
	g10	8	12	15	12	3	4	5	8	9	4
	regB	4	5	5	5	0	8	7	12	7	3

Table S3. Sequence read coverages of ligation products between the vector and the control genes. Sequence read amounts of toxic (*g10*, *g22*, *g38*, g137, *regB*) and non- toxic (*g119*, *g121*, *g150*, *g178*, *g246*) control genes from the ligation mixture and two replicate transformations. The sequences of the ligation joints in the predicted plasmids on both sides of the HPUF gene fragment were determined *in silico* and ca 15-25 nt of the plasmid and HPUF gene sequences of forward and reverse strands were extracted resulting in two pairs of complementary sequences for each cloning. Sequences named as gx_atg and gx_stop represent fragments from the leading strand containing start (atg) or stop (stop) codon of the gene insert. Gx_atgrev and gx_stoprev represents the corresponding fragments from the lagging strand. Elp1 and 2 stand for replicate transformations of the Ligation mix.

	Sequence	Ligation mix	Elp 1	Elp 2
*g10*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGATTAAGTACGATGT	36	11	29
*g10*_atgrev	TATACATCGTACTTAATCATGCGGCCGCCTGCAGGCATGCAAGCT	3	10	13
*g10*_stop	CAGCCTATGCTCACAGTAGCCATGGAGCTAGCTCTAGAGGATCCC	16	148	124
*g10*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGCTACTGTGAGCATAGGCTG	53	152	143
*g22*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGATTGACAGAGAAGA	193	22	13
*g22*_atgrev	TCTCTTCTCTGTCAATCATGCGGCCGCCTGCAGGCATGCAAGCTT	0	32	7
*g22*_stop	GGTGCGTGATGCATATTAACCATGGAGCTAGCTCTAGAGGATCCC	5	40	94
*g22*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTTAATATGCATCACGCACC	193	21	94
*g38*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGAAGTTAAAACACACTAG	1229	17191	16249
*g38*_atgrev	TTACTAGTGTGTTTTAACTTCATGCGGCCGCCTGCAGGCATGCAAGCT	3	14108	13946
*g38*_stop	GGGATCCTCTAGAGCTAGCTCCATGGTCATTCTGACCTCACTAAATG	614	17730	16471
*g38*_stoprev	CATTTAGTGAGGTCAGAATGACCATGGAGCTAGCTCTAGAGGATCCC	23	17760	16749
*g119*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGAAAACGTATAAAGAATT	877	11986	11896
*g119*_atgrev	AATTCTTTATACGTTTTCATGCGGCCGCCTGCAGGCATGCAAGCTTGG	4	11539	11460
*g119*_stop	TGGCACTAACGTTCGTTAACCATGGAGCTAGCTCTAGAGGATCCC	26	12216	12346
*g119*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTTAACGAACGTTAGTGCCA	567	11503	11787
*g121*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGAAAACCTATAATGAATT	156	9467	9211

*g121*_atgrev	AATTCATTATAGGTTTTCATGCGGCCGCCTGCAGGCATGCAAGCTTGG	2	10055	10311
*g121*_stop	GCTTAAAAAAGCTTCCTAACCATGGAGCTAGCTCTAGAGGATCCC	24	9823	9688
*g121*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTTAGGAAGCTTTTTTAAGC	170	9087	8714
*g137*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGAAAATTGCTGAACTAAT	16	7000	7362
*g137*_atgrev	ATTAGTTCAGCAATTTTCATGCGGCCGCCTGCAGGCATGCAAGCTTGG	0	8356	8866
*g137*_stop	ACAATTTCTAAGTCCTCATAGCCATGGAGCTAGCTCTAGAGGATCCC	15	9320	9287
*g137*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGCTATGAGGACTTAGAAATTGT	30	6716	6924
*g150*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGATTAAAGTTAATGAGC	1641	2398	2625
*g150*_atgrev	GCTCATTAACTTTAATCATGCGGCCGCCTGCAGGCATGCAAGCTTGG	3	2320	2489
*g150*_stop	CACGAATTGATATTGGATAGGCTAGCTCTAGAGGATCCCCGGGTAC	17	2162	2245
*g150*_stoprev	GTACCCGGGGATCCTCTAGAGCTAGCCTATCCAATATCAATTCGTG	1032	1950	2288
*g178*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGAGCAATATTAACCAGC	1387	7335	6952
*g178*_atgrev	GCTGGTTAATATTGCTCATGCGGCCGCCTGCAGGCATGCAAGCTTGG	0	6305	6202
*g178*_stop	AAACTAATAGCAGGATAACCATGGAGCTAGCTCTAGAGGATCCC	9	6111	5619
*g178*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTTATCCTGCTATTAGTTT	1342	6462	6341
*g246*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGTCTTTAAATGAAATG	1487	6475	6464
*g246*_atgrev	CATTTCATTTAAAGACATGCGGCCGCCTGCAGGCATGCAAGCTTGG	9	6219	6227
*g246*_stop	CATGCAAATGATTTTTAACCATGGAGCTAGCTCTAGAGGATCCC	19	6789	6478
*g246*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTTAAAAATCATTTGCATG	1396	5442	5254
*regB*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGACTATCAATACAG	0	0	0
*regB*_atgrev	CTGTATTGATAGTCATGCGGCCGCCTGCAGGCATGCAAGCTTGG	0	0	0
*regB*_stop	ATTAAAACTCAATGAGGTAAGGCTAGCTCTAGAGGATCCCC	0	0	0
*regB*_stoprev	GGGGATCCTCTAGAGCTAGCCTTACCTCATTGAGTTTTAAT	0	0	0

Table S4. Plating assay results of fHy-Eco03 HPUF genes.

	Control genes					Ligation 1					Ligation 2					Fraction of *g178*			
	CFU1	CFU2	CFU3	Average	SD	CFU1	CFU2	CFU3	Average	SD	CFU1	CFU2	CFU3	Average	SD	Lig 1	Lig 2	Average	SD
regB	11	9	17	12	3														
g178	391	451	447	430	27														
g01						353	321	419	364	41	186	303	175	221	58	0.847	0.515	0.681	0.166
g03						448	446	419	438	13	565	362	349	425	99	1.018	0.989	1.003	0.014
g04						263	312	247	274	28	278	324	311	304	19	0.637	0.708	0.672	0.035
g05						47	80	81	69	16	81	83	52	72	14	0.161	0.167	0.164	0.003
g06						400	389	343	377	25	225	222	185	211	18	0.878	0.49	0.684	0.194
g08						309	364	344	339	23	243	181	258	227	33	0.788	0.529	0.659	0.130
g09						432	400	423	418	13	312	290	277	293	14	0.973	0.681	0.827	0.146
g10						314	308	214	279	46	97	128	133	119	16	0.648	0.278	0.463	0.185
regB	9	12	9	10	1														
g178	417	282	314	338	58														
g11						399	380	388	389	8	173	149	337	220	84	1.151	0.65	0.900	0.250
g13						504	625	501	543	58	205	314	242	254	45	1.607	0.75	1.179	0.429
g15						482	479	513	491	15	244	282	239	255	19	1.454	0.754	1.104	0.350
g17						382	328	396	369	29						1.091		0.645	0.223
regB	7	8	12	9	2														
g178	397	415	413	408	8														
g23						196	191	205	197	6						0.484		0.682	0.199
g48						620	680	666	655	26						1.606		1.711	0.105
g50						449	370	382	400	35						0.981		0.747	0.234
g51						242	279	239	253	18	100	125	94	106	13	0.621	0.261	0.441	0.180
regB	11	10	10	10	0														
g178	742	597	794	711	83														
g52						486	348	331	388	69	335	319	252	302	36	0.546	0.425	0.485	0.061
g64						322	348	231	300	50	216	133	170	173	34	0.422	0.243	0.333	0.090
g65						293	307	200	267	47						0.375		0.514	0.139

g68						362	282	485	376	83	236	275	286	266	21	0.529	0.374	0.451	0.078
regB	6	7	8	7	1														
g178	384	372	377	378	5														
g69						430	399	340	390	46	213	234	186	211	24	1.031	0.558	0.795	0.236
g76						750	601	693	681	75	297	275	326	299	26	1.802	0.792	1.297	0.505
g77						496	590	565	550	49	190	209	158	186	26	1.456	0.491	0.974	0.482
g23											308	315	342	322	18		0.881		
g50											226	161	175	187	34		0.513		
g65											206	245	265	239	30		0.654		
g17											216	255	235	235	20		0.645		
g48											650	660	679	663	15		1.816		
regB	14	7	14	12	3														
g178	381	314	399	365	37														

Table S5. Sequence read coverages of ligation products between the vector and the fHy-Eco03 HPUF genes. The sequences of the ligation joints in the predicted plasmids on both sides of the HPUF gene fragment were determined *in silico* and ca 15-25 nt of the plasmid and HPUF gene sequences of forward and reverse strands were extracted resulting in two pairs of complementary sequences for each cloning. Sequences named as gx_atg and gx_stop represent fragments from the leading strand containing start (atg) or stop (stop) codon of the gene insert. Gx_atgrev and gx_stoprev represents the corresponding fragments from the lagging strand. Lig1Elp1/2 stand for replicate transformations of Ligation mix 1 and Lig2Elp1/2 stand for replicate transformations of Ligation mix 2.

	Sequence	Ligation mix 1	Ligation mix 2	Lig1Elp1	Lig1Elp2	Lig2Elp1	Lig2Elp2
*g77*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGAAAACAATGTTTGTATACAAG	1426	870	5061	3957	5141	4112
*g77*_atgrev	CTTGTATACAAACATTGTTTTCATGCGGCCGCCTGCAGGCATGCAAGCTTGG	12	6	3902	3272	4273	3045
*g77*_stop	GAGGTGTTAGGCCTGTGACCATGGAGCTAGCTCTAGAGGATCCC	49	35	4599	3988	5007	4006
*g77*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTCACAGGCCTAACACCTC	709	681	4720	4073	5396	4310
*g76*_atg	AGCTTGCATGCCTGCAGGCGGCCGCATGAAACATCAGGGC	355	302	5815	4711	4355	3411
*g76*_atgrev	GCCCTGATGTTTCATGCGGCCGCCTGCAGGCATGCAAGCT	1	2	4961	4305	3949	3037
*g76*_stop	CATGCACGCTGGAGATAACCATGGAGCTAGCTCTAGAGGATCCC	40	26	4816	3867	3986	2842
*g76*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTTATCTCCAGCGTGCATG	353	226	5026	4241	3925	3084
*g69*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGGAAGCCACCAAGTCT	965	556	6676	5661	5057	4117
*g69*_atgrev	AGACTTGGTGGCTTCCATGCGGCCGCCTGCAGGCATGCAAGCTTGG	3	2	6311	5534	4816	3859
*g69*_stop	GTGTACCATTGAAGTGTATAAATAACCATGGAGCTAGCTCTAGAGGATCCC	19	11	5559	4802	4518	3431
*g69*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTTATTTATACACTTCAATGGTACAC	687	467	6244	5494	5048	4024
*g68*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGTATACGTTCAATCAATTCAAA	293	205	1348	1086	846	678
*g68*_atgrev	TTTGAATTGATTGAACGTATACATGCGGCCGCCTGCAGGCATGCAAGCTTGG	8	7	1157	1034	757	553
*g68*_stop	GATGTTAACAGAGGGGTATTAACCATGGAGCTAGCTCTAGAGGATCCC	42	42	1313	1020	936	678
*g68*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTTAATACCCCTCTGTTAACATC	336	229	1305	1039	848	732
*g65*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGCAGAACCTTTGGGATA	533	444	5013	4535	5843	4669
*g65*_atgrev	TATCCCAAAGGTTCTGCATGCGGCCGCCTGCAGGCATGCAAGCTTGG	0	3	4243	3836	5056	3961
*g65*_stop	CCGGCTACACAAGCTGACCATGGAGCTAGCTCTAGAGGATCCC	27	30	4708	4347	5995	4460
*g65*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTCAGCTTGTGTAGCCGG	380	369	5268	4669	6475	4800

*g64*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGTTTGTGCCGATTCCG	837	497	2055	2072	1918	1371
*g64*_atgrev	CGGAATCGGCACAAACATGCGGCCGCCTGCAGGCATGCAAGCTTGG	9	4	1933	1847	1814	1396
*g64*_stop	CGCTGAAGAAGCGAATTGACCATGGAGCTAGCTCTAGAGGATCCC	30	17	2068	2012	1976	1520
*g64*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTCAATTCGCTTCTTCAGCG	746	479	2320	1994	2049	1575
*g52*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCGTGTCTAATATAGGTTTAGAGTACC	562	502	4254	3525	3656	2843
*g52*_atgrev	GGTACTCTAAACCTATATTAGACACGCGGCCGCCTGCAGGCATGCAAGCTTGG	6	19	3384	2739	3070	2244
*g52*_stop	CGAGGAAGAATGCAAATGACCATGGAGCTAGCTCTAGAGGATCCC	35	40	4448	3531	3906	3292
*g52*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTCATTTGCATTCTTCCTCG	606	465	4127	3439	3598	2899
*g51*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGAGAAAGTTTACCCTGGA	50	42	5333	4457	4895	3957
*g51*_atgrev	TCCAGGGTAAACTTTCTCATGCGGCCGCCTGCAGGCATGCAAGCTTGG	4	4	6204	5176	5605	4309
*g51*_stop	ACTTGGCTAAGAAATATGATTGACCATGGAGCTAGCTCTAGAGGATCCC	27	9	5561	4843	5635	4481
*g51*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTCAATCATATTTCTTAGCCAAGT	85	80	5020	4254	4697	3650
*g50*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGATTGAATCTACACGCAC	244	162	5148	4293	3726	2957
*g50*_atgrev	GTGCGTGTAGATTCAATCATGCGGCCGCCTGCAGGCATGCAAGCTTGG	5	3	4566	3610	2966	2280
*g50*_stop	AAAAGAAAGGAAAGCGCTAACCATGGAGCTAGCTCTAGAGGATCCC	40	41	4639	3793	3206	2737
*g50*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTTAGCGCTTTCCTTTCTTTT	233	155	4692	3994	3447	2807
*g48*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCTTGACCTCATGCAGTAAAAAG	241	182	16372	14621	14149	11134
*g48*_atgrev	CTTTTTACTGCATGAGGTCAAGCGGCCGCCTGCAGGCATGCAAGCTTGG	6	3	14949	13359	12842	9941
*g48*_stop	AGTGAAGGGGCCAATTAACCATGGAGCTAGCTCTAGAGGATCCC	87	79	15501	14418	14516	11213
*g48*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTTAATTGGCCCCTTCACT	282	247	16598	15055	15410	11945
*g23*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGGGTGGCGCAAGAATC	689	461	3220	2952	2936	2134
*g23*_atgrev	GATTCTTGCGCCACCCATGCGGCCGCCTGCAGGCATGCAAGCTTGG	4	4	3019	2658	2716	2228
*g23*_stop	TTTACTTGTACAAAATTAGCCATGGAGCTAGCTCTAGAG	68	44	3380	3204	3293	2587
*g23*_stoprev	CTCTAGAGCTAGCTCCATGGCTAATTTTGTACAAGTAAA	1522	1213	3655	3378	3346	2680
*g17*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGGACTTATTCGACTTAGTTGA	639	455	6207	5269	4906	4122
*g17*_atgrev	TCAACTAAGTCGAATAAGTCCATGCGGCCGCCTGCAGGCATGCAAGCTTGG	8	2	5034	4373	4366	3212
*g17*_stop	TGCTGGTAGAAATGGCATAACCATGGAGCTAGCTCTAGAGGATCCC	65	46	6295	5211	5472	4377
*g17*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTTATGCCATTTCTACCAGCA	809	555	5995	5116	5124	4126
*g15*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGAACATCATAGTTGAGGGA	820	595	6502	5798	3773	3065
*g15*_atgrev	TCCCTCAACTATGATGTTCATGCGGCCGCCTGCAGGCATGCAAGCTTGG	3	5	6322	5494	3775	2777

*g15*_stop	CGATAAGGAGTGGACTAAATGACCATGGAGCTAGCTCTAGAGGATCCC	42	29	6070	5806	3783	3228
*g15*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTCATTTAGTCCACTCCTTATCG	881	488	6101	5753	3711	3146
*g13*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGACTGTAGTTGCACCT	183	250	6229	5350	7414	5717
*g13*_atgrev	AGGTGCAACTACAGTCATGCGGCCGCCTGCAGGCATGCAAGCTTGG	2	7	4250	3905	5337	3939
*g13*_stop	TGGCTTAAATTTAAGGTAAAATGACCATGGAGCTAGCTCTAGAGGATCCC	12	21	4731	4310	6048	4987
*g13*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTCATTTTACCTTAAATTTAAGCCA	224	251	4921	4266	6056	4915
*g11*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGCATACTCAATTAATCATCTGG	71	67	2949	2514	2725	2207
*g11*_atgrev	CCAGATGATTAATTGAGTATGCATGCGGCCGCCTGCAGGCATGCAAGCTTGG	4	5	2945	2530	2821	2149
*g11*_stop	CACAATTTAACTCAGGAGAATGACCATGGAGCTAGCTCTAGAGGATCCC	29	20	3157	2815	3035	2447
*g11*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTCATTCTCCTGAGTTAAATTGTG	90	68	3006	2491	2800	2282
*g10*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGATGAAAGAATTGACATTGAC	80	28	2089	1723	1148	916
*g10*_atgrev	GTCAATGTCAATTCTTTCATCATGCGGCCGCCTGCAGGCATGCAAGCTTGG	5	60	2152	1714	1290	943
*g10*_stop	TCAAATTAGTACAAAGTTACAATAGCCATGGAGCTAGCTCTAGAGGATCCC	23	11	2273	1595	1317	985
*g10*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGCTATTGTAACTTTGTACTAATTTGA	92	60	1807	1549	1120	846
*g09*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGGTAACATTCACAACTATCC	251	142	5079	4545	3602	3039
*g09*_atgrev	GGATAGTTGTGAATGTTACCATGCGGCCGCCTGCAGGCATGCAAGCTTGG	10	3	5054	4295	3427	2766
*g09*_stop	GACGAATCAAAATTTGAGGTTTAACCATGGAGCTAGCTCTAGAGGATCCC	35	29	4920	4132	3430	2741
*g09*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTTAAACCTCAAATTTTGATTCGTC	281	178	4617	4191	3383	2828
*g08*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGCCATTAATCAAAGTTACAG	80	46	4148	3417	3805	3052
*g08*_atgrev	CTGTAACTTTGATTAATGGCATGCGGCCGCCTGCAGGCATGCAAGCTTGG	3	1	3528	3075	3626	2631
*g08*_stop	GTTACTGAGGTAATTATTTCTTGACCATGGAGCTAGCTCTAGAGGATCCC	41	26	4010	3509	4102	3220
*g08*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTCAAGAAATAATTACCTCAGTAAC	121	106	3581	3219	3749	2867
*g06*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGAAAAAGCATGTTGTTGAG	27	26	3666	3012	3342	2638
*g06*_atgrev	CTCAACAACATGCTTTTTCATGCGGCCGCCTGCAGGCATGCAAGCTTGG	8	11	3637	3076	3264	2626
*g06*_stop	GGTGTTGTCTTATGTTTTCTGACCATGGAGCTAGCTCTAGAGGATCCC	156	175	3611	3256	3578	2895
*g06*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTCAGAAAACATAAGACAACACC	46	36	3562	3072	3373	2691
*g05*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGTTTTCTGAGGAGCAACT	283	178	0	3	4	0
*g05*_atgrev	AGTTGCTCCTCAGAAAACATGCGGCCGCCTGCAGGCATGCAAGCTTGG	2	1	0	0	0	0
*g05*_stop	ACTTGGAGTAAGGAGCAAATAACCATGGAGCTAGCTCTAGAGGATCCC	42	25	21	10	12	8
*g05*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTTATTTGCTCCTTACTCCAAGT	335	248	33	38	5	12

*g04*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGTTTAAATTCTTACAACGTAACC	58	53	3239	2652	2607	2197
*g04*_atgrev	GGTTACGTTGTAAGAATTTAAACATGCGGCCGCCTGCAGGCATGCAAGCTTGG	4	2	3014	2477	2352	1885
*g04*_stop	GTTGAAGGAGATTACTTGTGACCATGGAGCTAGCTCTAGAGGATCCC	26	18	3240	2553	2760	2184
*g04*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTCACAAGTAATCTCCTTCAAC	77	56	3347	2763	2729	2243
*g03*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCGTGATTGGCTTAATCCTGG	104	81	5920	5084	4659	3780
*g03*_atgrev	CCAGGATTAAGCCAATCACGCGGCCGCCTGCAGGCATGCAAGCTTGG	2	2	5973	4799	4627	3435
*g03*_stop	CGATGAGTTGAAACAAGACTAACCATGGAGCTAGCTCTAGAGGATCCC	14	24	5938	5086	4863	3754
*g03*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTTAGTCTTGTTTCAACTCATCG	103	88	5770	4927	4656	3613
*g01*_atg	CCAAGCTTGCATGCCTGCAGGCGGCCGCATGAGCTATACTGACCAACA	95	55	3562	3244	3162	2449
*g01*_atgrev	TGTTGGTCAGTATAGCTCATGCGGCCGCCTGCAGGCATGCAAGCTTGG	0	2	3930	3443	3243	2451
*g01*_stop	CCTGAGAGGTCAACTGTAACCATGGAGCTAGCTCTAGAGGATCCC	24	27	3704	3186	3293	2471
*g01*_stoprev	GGGATCCTCTAGAGCTAGCTCCATGGTTACAGTTGACCTCTCAGG	126	105	3665	3312	3257	2484

S6. Workflow of the NGS assay bioinformatics

The raw NGS read data was analyzed using the Puhti computer environment at CSC (the Finnish Centre for Scientific Computing) following the protocol outlined below.

1. The sequences over the ligation joints in the predicted plasmids on both sides of the HPUF gene fragment were determined *in silico* as described for Table S5 and illustrated in Figure 2B. A text document containing each sequence on its own line was prepared and saved under the name **list.txt**. This file was uploaded by WinSCP to the Puhti directory containing the NGS raw data files.
2. Bio tools were activated using the commands

```
$ module load biokit
```
and
```
$ module load velvet
```

3. The compressed fastq.gz NGS sequence read files were uncompressed with the **gunzip** command

```
$ gunzip file_name.gz
```

4. The paired end fastq-files were interleaved to a single file using the **shufflesequences** command.

```
$ shuffleSequences_fastq.pl read_file_1.fq read_file_2.fq file_name.fastq
```

5. Alignments were run as a batch job by first editing the bash script text file template (file_name.sh) to include the required file names and paths.

```
#!/bin/bash -l
#SBATCH -o std1.out
#SBATCH -e std1.err
#SBATCH -p small
#SBATCH --account=skurnik
#SBATCH --ntasks=1
```

```
#SBATCH --cpus-per-task=1
#SBATCH --nodes=1
#SBATCH -t 48:00:00
#SBATCH --mem=128000

module load biokit
#change directory to the one where you have the data
cd /file_path

# insert the file name for your sequence fragment text file
a=1
for pat in $(cat sequencelist_file_name.txt)
do
#insert the fastq filename of your interleaved paired end reads
  fuzznuc -pattern "$pat" file_name.fastq -rformat excel -filter | awk '{ if ( $1 != "SeqName") print $1}' | sort | uniq > name_$a
  (( a = a + 1))
done
```

6. The bash file can be edited using Nano (version 2.3.1) with the **nano** command

```
$ nano file_name.sh
```

7. To save the changes and to exit Nano the CTRL+O and CTRL+X commands were used
8. The batch job was submitted by using the **sbatch** command

```
$ sbatch file_name.sh
```

9. The batch job results in a number of files with a name containing a running number corresponding to each sequence line in the **list.txt** file created in point 1. Each name_N file created contains the information of the number of reads containing searched sequence.